Crowdsourced Intuitive Visual Design Feedback

David Allan Robb

Thesis submitted for the

Degree of Doctor of Philosophy

Heriot-Watt University

School of Mathematical and Computer Sciences

May 2015

Abstract

For many people images are a medium preferable to text and yet, with the exception of star ratings, most formats for conventional computer mediated feedback focus on text. This thesis develops a new method of crowd feedback for designers based on images. Visual summaries are generated from a crowd's feedback images chosen in response to a design. The summaries provide the designer with impressionistic and inspiring visual feedback. The thesis sets out the motivation for this new method, describes the development of perceptually organised image sets and a summarisation algorithm to implement it. Evaluation studies are reported which, through a mixed methods approach, provide evidence of the validity and potential of the new image-based feedback method.

It is concluded that the visual feedback method would be more appealing than text for that section of the population who may be of a visual cognitive style. Indeed the evaluation studies are evidence that such users believe images are as good as text when communicating their emotional reaction about a design. Designer participants reported being inspired by the visual feedback where, comparably, they were not inspired by text. They also reported that the feedback can represent the perceived mood in their designs, and that they would be enthusiastic users of a service offering this new form of visual design feedback.

Acknowledgements

"No man is an island" (Donne 1623). I thank and acknowledge the following:

- My supervisors, Professor Mike Chantler and Dr Britta Kalkreuter, for their sage direction and for the idea of using images as a medium for crowdsourced feedback. That idea was the premise of the project that I was recruited in 2011 to pursue.
- Heriot-Watt Creativity, Design and Innovation Theme as project funder.
- Dr Fraser Halley on whose shoulders that part of my work involving perceptual similarity enriched by crowds, the SOM browser, and 3D visualisations of perceptually organised image sets, stands. His provision of code exemplars allowed the project to hit the ground running (see Bibliography).
- Dr Stefano Padilla whose judgement I sought at many stages of the project.
- My wife, Mrs Alexandra Robb, for proofreading and foregoing the benefits of having a husband with a well-paid job while I was engaged on my postgraduate work.
- Dr Patrick Green (now retired as a Professor from the Psychology Department at Heriot-Watt University) for advice on one of my experiment designs.
- Dr B. Kalkreuter for administering the design emotion terms survey (in Chapter 8).
- Dr Andrew Lynn of Strathclyde University for his help in placing HITs on Amazon Mechanical Turk (in Chapter 4).
- My Texture Lab colleagues for their cheerful advice and encouragement.
- My post-graduate colleague, Mrs Chamithri Greru, for admin support during the main evaluation experiment.
- My daughter, Miss Jane Robb, for proofreading.
- And finally, my experiment participants, both known and anonymously crowdsourced, who took part in this work's numerous experimental sessions. (See Appendix G p.240 for details on the numbers).

Publications

Here the publications associated with the thesis are listed and the contribution of the author of this thesis is made explicit. First the contribution of co-authors Padilla, Kalkreuter and Chantler thematically across all the publications is stated. Then the specific contribution of Padilla in relation to the experiment described in Chapter 7 (and described in a subset of the papers) is stated. Finally, the publications are listed in order of weight of the author's contribution (where this contribution is equal, papers are listed in reverse chronological order). With each paper citation there is a description making explicit the contribution of the Author of this thesis.

Co-author contributions across the papers (Padilla, Kalkreuter and Chantler)

Some aspects of the project feature in all of the papers. Specifically, the Crowdsourced Visual Feedback Method (CVFM), its possible benefits to the design process, and the use of perceptual data for image summarisation. Co-authors Padilla, Kalkreuter and Chantler contributed the idea of the CVFM (described in Chapter 1), and of using perceptual data as the basis for summarisation. Please also note their mentions in the Acknowledgements.

Padilla and the work of Chapter 7

Padilla made a practical contribution to the work of Chapter 7 by providing a previously constructed experiment application (Padilla, 2011) used for Task 1 of the experiment described diagrammatically in Figure 7.1, and detailed explicitly in Section 7.2.1. That application presented a list of terms as stimuli for subjects to select images. That list of terms was researched and compiled by Padilla. Where a paper described below describes the work of Chapter 7 then Padilla's contribution is correspondingly both an intellectual and a practical contribution.

The papers

1. Robb, D.A., Padilla, S., Kalkreuter, B., & Chantler, M.J. (2015b). Moodsource: Enabling Perceptual and Emotional Feedback from Crowds. *CSCW'15: Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, Companion*. pp. 21-24.

Written by the Author. All the work described is that of the Author.

2. Robb, D.A., Padilla, S., Kalkreuter, B., & Chantler, M.J. (2015a). Crowdsourced Feedback With Imagery Rather Than Text: Would Designers Use It? *CHI'15: Proceedings of the 33rd ACM Conference on Human Factors in Computing Systems*. pp.1355-1364.

   Written by the Author. (The paper is an iteration of two prior submissions in the first of which Chantler contributed as editor).

3. Kalkreuter, B., Robb, D., Padilla, S., & Chantler, M.J. (2013), Managing creative conversations between designers and consumers, in *Britt, H., Wade, S., Walton, K. (Eds), Futurescan 2: Collective Voices,* Association of Fashion and Textile Courses, Sheffield, 90-99.

   Edited by Kalkreuter. Kalkreuter wrote the design context and most of the semiological context parts of the paper. The Author wrote the parts describing the practical and experimental work and provided the figures. (This is the experiment described in Chapter 7).

4. Kalkreuter, B., & Robb, D. (2012). HeadCrowd: Visual feedback for design. *Nordic Textile Journal, Sustainability and Innovation in the Fashion Field* (1), 70-81.

   Edited by Kalkreuter. The Author contributed the literature review. The paper focusses on the early output from the experimental Task 1 described in Chapter 7. The Author contributed the description of the construction of the abstract image set, visualisations of the task image outputs, and provided supporting figures. Kalkreuter contributed the overall context for the paper and the specific context for the experiment. Kalkreuter also contributed the quantitative analysis of the results and the conclusions.

5. Padilla, S., Robb, D., Halley, F., & Chantler, M.J. (2012). Browsing Abstract Art by Appearance. *Predicting Perceptions: Proceedings of the 3rd International Conference on Appearance*, 100-103.

   Written by Padilla. The Author contributed the image set described in the paper (see Chapter 4) and one of the figures (i.e. that showing a 5x5 stack SOM browser presentation of the image set).

6. Padilla, S., Halley, F., Robb, D., & Chantler, M., (2013), Intuitive Large Image Database Browsing Using Perceptual Similarity Enriched by Crowds, *Computer Analysis of Images and Patterns*, *LNCS*, Springer, 8048, 169-176

   Padilla did the experimentation and wrote the paper which also draws on the work of Halley. The Author's contribution is in the form of one of the two perceptually organised image sets described in the paper, specifically the abstract image set. (See Chapter 4).

Contents

List of Figures

List of Tables

Terminology

Term	Definition
Abstracts	The abstract images answer format when the context is feedback participants; The abstract images visual summary feedback format when the context is designer participants.
Abstract500	The set of 500 images of a loosely abstract nature used to populate a SOM browser for feedback use.
AMT	Amazon Mechanical Turk, a crowdsourcing tool
CBIR	Content Based Image Retrieval
Crowdsourced visual feedback method	The proposed method of feedback being developed in this thesis and illustrated in Figure 1.1. (See CVFM)
CVFM	Crowdsourced visual feedback method (see above)
Design feedback emotion subset	A subset of 19 terms from the Plutchik (2003) emotion model suitable for design feedback.
ECI	The emotion categoriser for images. The application constructed to allow unsupervised image emotion categorisation by drag and drop.
Emotion profile	The emotion (category frequency) profile of an image in the Emotive2000 is the pattern of emotions associated with it as described by its tag frequency vector and/or its term frequency vector.
Emotives	The emotive images answer format when the context is feedback participants; The emotive images visual summary feedback format when the context is designer participants.
Emotive204	A subset of the Emotive 2000 filtered by emotion profile (specifically, the term frequency vector) to produce a set of images balanced across the 19 terms of the design feedback emotion subset.
Emotive2000	The full set of Creative Commons images for which emotion profiles were gathered.
f-1st	Frequency of first rank for intended meaning. The frequency with which participants ranked a visual stimulus' intended meaning first among other meanings.
FPR-Theme	Themes arising out of the reasons stated by designer participants for their format preference rankings during interview.
HIT	Human Intelligence Task (Kazai 2011). One task done for payment by a participant via a crowdsourcing tool such Amazon Mechanical Turk (AMT).
HWU	Heriot Watt University
ISL	Image selection list. A list of image IDs selected from a given image set by a group of participants to communicate some idea. The list can contain repetitions of any image ID and thus reflects the popularity of that image within the selection.
MDS	Multidimensional scaling
Response format	During evaluation of the feedback method three response formats were used by feedback participants. These were 1) choose 3 images from the abstract image browser 2) choose 3 images from the emotive image browser and 3) enter text.
QC	Quality Control. Mainly concerned with the production and use of gold data to establish the reliability of crowdsourced observations.

Term	Definition
SOM	Self-organising map. A low-dimensional representation of multidimensional data produced through an unsupervised learning algorithm. Originally conceived for neural networks (Kohonen 1990).
Tag frequency vector	A 56-member vector, each element of which describes the frequency with which an image was tagged as belonging on given numbered spot or tag location on the Plutchik emotion model (Figure 8.2).
Tag location	One of the 56 numbered spots or tag locations on the Plutchik emotion model (Figure 8.2) used for tagging emotive images.
Term frequency vector	A 32-member vector each element of which describes the frequency with which an image was tagged by dropping it on the Plutchik emotion model on spots corresponding to a given emotion term, e.g. love.
TEX	Heriot Watt University School of Textiles and Design which has a Campus in the Scottish Borders.
Valence	One of the emotion dimensions: The intrinsic attractiveness or repulsiveness of a stimulus. E.g. an image of a happy smiling child would evoke high (positive) valence, whereas an image of bodily mutilation would evoke low (negative) valence.
VAS	Visual Analogue Scale: A self-report response method in which respondents place a cross on a linear scale anchored on either end by semantically opposite terms e.g. Hot ---------- Cold. A VAS item consists of a question and its accompanying anchored response scale.
Vector	This term is used in its C+ or MATLAB programming sense: A sequence (or one-dimensional array) of numbers, each member of which describes an aspect or feature of some entity.

Chapter 1

Introduction

1.1 Motivation and Statement of the Problem

Design is an important activity economically. While in general it forms a small portion of the cost of products (three to seven percent), around 80 percent of manufacturing costs are expended during the first 20 percent of the design process (Goel & Pirolli, 1992). Thus mistakes during the design process can be costly for individual projects or products. Design, as a cognitive activity, faces a particular challenge. It has both logical and creative aspects, and these require different abilities in the designer (Archer, 1969). Thus designers bear a burden of responsibility to get it right to avoid waste of resources whilst facing a cognitive challenge.

Feedback is an important aspect of the design process and can help designers iterate towards an optimal solution. While the experiments in this thesis focus on the domains of fashion and interior design, these share much in common with other aesthetic design domains such as graphic and product design in, for example, automotive, food, and travel industries. Designers, particularly in aesthetic domains such as these, face an asymmetry in terms of their design medium on one side and the medium of conventional forms of feedback that they might expect to receive on the other. Their design medium and indeed much of their inspiration is largely visual (Garner & McDonagh-Philp, 2001) whereas any feedback they might receive, be it locally or remotely from peers, or other domain experts, is usually textual or verbal. Also those involved in giving design feedback tend to be connected professionally to design, or part of an enthusiastic interest group and can be, to an extent, disconnected from the end users of the product being designed (Cook et al., 2009) (Xu et al., 2014).

The emotional and intuitive reactions of potential consumers or interlocutors to a given product or idea are an important factor in whether or not that product or idea will become popular (Taylor, 2000). There are psychological reasons why images can often

be a more effective medium than text for communication, particularly where emotion is involved (Riding & Ashmore, 1980) (Junghöfer et al., 2001).

Irrespective of whether or not a given individual is a potential consumer of a product or idea their judgement on it can be of value when they form part of a crowd; the collective judgement of a crowd of non-experts, can often be as accurate as that of an expert in a given domain (Surowiecki, 2008). When the subject of feedback is a prototype design there is potential for a beneficial cycle of feedback and refinements to the design analogous to a conversation during which the design develops.

Statement of the problem

This thesis proposes that there is a need for a method of engaging crowds in visual feedback for designs to a) allow designers to connect with potential users of their products in a less formal way than conventional methods allow and b) give designers access to crowd feedback which is less structured and more visual, intuitive, and potentially inspiring than current modes of crowd design feedback.

Further background to the motivation is discussed in further detail in Chapter 2.

This chapter continues with a description of the thesis goals which include the development of a method of obtaining crowdsourced visual feedback in Section 1.2. Then, in 1.3, the thesis scope is defined. Section 1.4 describes the original contributions from this work and, lastly, Section 1.5 sets out the thesis organisation.

1.2 Goals

Summarising the motivation set out in 1.1: the role of feedback in the design process, the value of a body of non-expert opinion, the importance of emotional and intuitive reactions to a product or idea, the predominance of text as the medium for computer mediated feedback, and the potential benefits of enabling crowd participation in the design process have prompted this proposition of a method of crowdsourced design feedback based on images. Figure 1.1 illustrates the proposed method and where the main contributions of this thesis lie. Originally the idea was motivated by fashion designers and in this thesis it is evaluated with fashion and interior design students. It is expected to apply to any aesthetic design context including product and graphic design, in automotive, food, travel and other sectors.

The goals of this thesis are a) to develop the means to implement this method of crowdsourced visual feedback sufficiently to allow its evaluation, and then b) to evaluate it.

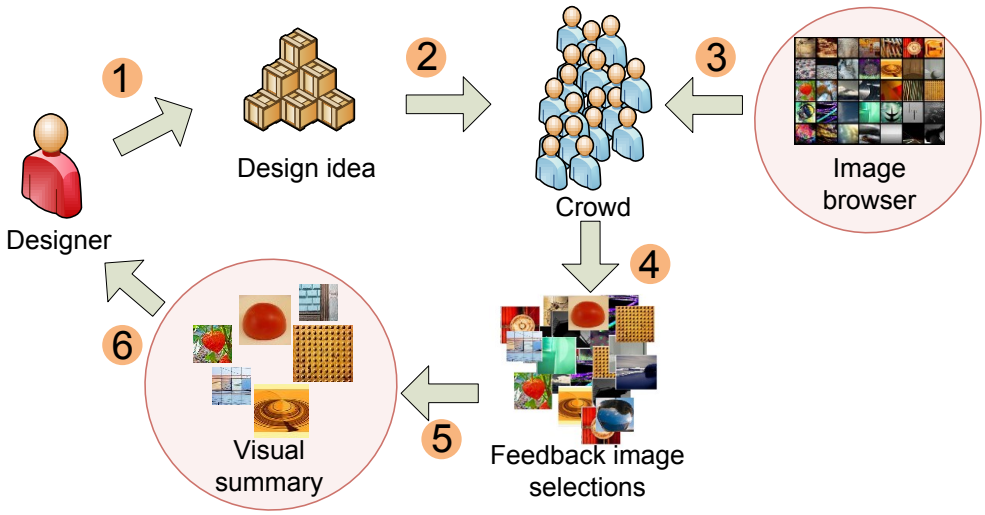


Figure 1.1 - The crowdsourced visual feedback method (CVFM) highlighting the areas of novelty: 1-Designer presents design; 2-Crowd views design; 3- Crowd responds by selecting images from browser; 4- Image selections collated; 5-Visual summary generated; 6-Designer views feedback summary. Circles highlight where the major contributions of this thesis lie: images as the medium for feedback; image summarisation applied to crowd communication; an evaluation of the communicative effectiveness of aggregated image selections relative to summaries; an evaluation of the CVFM with groups of designer users and crowd users.

In the rest of this thesis, *crowdsourced visual feedback method* is abbreviated to CVFM.

1.3 Scope

While the ideas behind the project encompass a number of areas including semiology, participatory design, design feedback, intuition, visual communication, the communication of emotion, marketing, sustainable consumption, social computing, and crowdsourcing, the scope of this thesis must, by necessity, be limited. Those areas listed are discussed in the chapters of this thesis where appropriate to introduce and provide context for the focus of the thesis which is described in 1.3.1. Exclusions are set out in 1.3.2.

1.3.1 Included in this Thesis

The following are dealt with:

a) The development of a method of computer mediated communication for involving crowds in design feedback through the medium of images.

b) The construction of a perceptually organised image browser populated with abstract images with which to investigate the method.

c) The development and implementation of an algorithm for summarising massed image selections from the abstract image browser.

d) An evaluation of the communicative effectiveness of selections of images chosen from the abstract image browser compared to summaries of those selections formed using the summarisation algorithm.

e) The development of a further image set with the aim of communicating emotions better than can be done with the abstract image set.

f) Two evaluation studies (one a pilot of the second) of a single cycle of feedback communication using the CVFM with two groups of students at a design institution. In both studies a small number of students took part as designers, putting forward their designs, while the remainder participated as the feedback crowd.

1.3.2 Exclusions

Part of the motivation to develop crowdsourced design feedback using images is to exploit and access human intuition and cater for aspects of cognition. However, this thesis does not set out to specifically measure psychological traits (such as visual and verbal cognitive styles) within participants taking part in the evaluations. This aspect is discussed as a possible direction in future work.

The CVFM (Figure 1.1) is evaluated using prototypes. An end-to-end, operational, service to implement the method, with cycles of feedback, is not developed.

The possibility of using the CVFM in cycles of prototype presentation, initial feedback, prototype modification, and further feedback, is discussed but subsequent cycles of feedback (beyond the first) are not evaluated by studies in this thesis.

Ultimately the CVFM requires a crowd or crowds to be motivated to engage with it. While the appeal for feedback users (the crowd) is evaluated during this thesis, the actual mechanism by which a designer user would connect with a crowd or recruit their own crowd is not investigated beyond being discussed speculatively.

1.4 Original Contributions

The contributions of this thesis fall into three categories: new methods, data sets, and evaluation studies.

1.4.1 New Methods or New Application of Existing Methods

The development of a method of crowdsourced visual feedback based on images

Referring to Figure 1.1, systems offering a process of computer mediated design feedback do already exist, encompassing designers presenting to a crowd and receiving some processed feedback. However, the involvement of an image browser as a medium for the feedback from a crowd is novel and is not a feature of any existing system. The application of visual summarisation to the image selections of a crowd is also novel and has not been achieved prior to this work.

Use of perceptual data to summarise the image choices of a crowd

The use of purely perceptual data gathered on large (>200) collections of images *for the purpose of summarisation* is also new.

The construction of an emotion image set where each image is described by a crowdsourced emotion category frequency vector

This had not been done before. (Although texture property labelling frequency has just recently been used to study the automated extraction of such semantic properties from a set of texture images (Cimpoi et al., 2014).)

1.4.2 Data Sets

During the work for this thesis two sets of Creative Commons images were assembled and perceptual data gathered on them. Both image sets along with a) their perceptual data and b) their attribution data (required by the Creative Commons licences) are a) in

the Thesis Additional Materials optical disk (see 1.5) and b) available for download from the Heriot Watt University Texture Lab web site[1] (in the Resources section). They will be accessed via a page whose content when indexed by search engines will allow the location of these resources by search should they at some time in the future be moved.

The next two subsections describe each image set and its specific copyright status.

Abstract image set

This is a set of 500 images of a mainly abstract nature (*Abstract500*) and its accompanying perceptual similarity matrix.

Each image downloaded for this image set from Flickr had an Attribution-Non Commercial-Share Alike 2.0 Generic (CC BY-NC-SA 2.0) licence. This permits sharing and adaptation. Adaptation was required as many of the images were cropped to a square aspect ratio.

Emotive image set

This consists of 2000 images (*Emotive2000*) and its accompanying emotion categorisation frequency data.

Each image downloaded had, as a minimum, an Attribution-Non Commercial-NoDerivs 2.0 Generic (CC BY-NC-ND 2.0) licence. This permits sharing. Only sharing was required as resizing was the only modification made to the images. The licence permits resizing.

1.4.3 Evaluation Studies

Two evaluation studies are reported in this thesis.

Chapter 7

This study evaluates the communicative effectiveness of the Abstract500 for material and emotional terms. The same study allowed the evaluation of the communicative effectiveness of the novel visual summaries.

---

[1] http://www.macs.hw.ac.uk/texturelab

Chapters 9 and 10

The other study (Chapter 10) evaluated the CVFM with a group of undergraduate students including interior designers. A pilot was carried out for this, using student fashion designers (Chapter 9).

1.5 Thesis Organization

This thesis is linear in organisation. In line with the thesis goals, *"a) to develop the means to implement the CVFM sufficiently to allow it to be evaluated*, *and then b) to evaluate it."* the chapters form two groups. The first group, Chapters 2 to 8, establish the motivation for the CVFM and develop the means to enable the evaluation. Then the second group, Chapters 9 to 11, describe the evaluation and conclusions.

Additionally, the chapters fall into three categories:

- Theoretical (or literature based).
- Practical: a component for the method is developed.
- Evaluative: some aspect or the whole of the method is evaluated either experimentally or by discussion.

Where required, a chapter has a single appendix associated with it. The appendix name is noted following each chapter introduction if required. Reference to part of an appendix is done by appendix name and a page number. The appendices perform two functions. They contain detailed results where these require reporting but the appendices also hold details of materials and processes included to allow another researcher to replicate the work by referring to the thesis, appendices and to *Additional Material* (on an accompanying optical disk). The Additional Material contains data sets, application code, scripts, input and output files, database tables and records of procedures followed. It is structured such that the material can be located by chapter title and appendix sub-heading.

Where the work described in a chapter features in a publication, that fact is noted in the chapter introduction under a heading "Published Work". Such notes briefly refer to the publications concerned.

The agenda of each chapter is briefly described below, firstly the development chapters and secondly the evaluation chapters.

1.5.1 Development Chapters

Chapter 2 (Theoretical) details the background to the motivation in developing the CVFM to encourage participatory design, to improve on conventional methods of feedback, and to use images as the medium.

Chapter 3 (Theoretical) describes the background to the selection of the perceptually organised self-organising map (SOM) browser as a component of the implementation.

Chapter 4 (Practical) relates the construction of a SOM browser populated with abstract images (Abstract500) from which the crowd can select images as their feedback response.

Chapter 5 (Theoretical) describes a) the necessity for image summarisation, b) existing work in the area and c) argues for a particular approach involving clustering based on perceptual data as suitable in this case.

Chapter 6 (Practical) describes the development of an algorithm for image summarisation as a component in the implementation of the CVFM. The summarisation allows multiple image selections from the Abstract500 to be summarised in a few representative images.

Chapter 7 (Evaluative) describes a study evaluating and comparing the communicative effectiveness of image selections from the Abstract500 browser and visual summaries that were generated from them.

Chapter 8 (Practical) details the construction of a second image browser populated with images more suited to communicating emotion than the Abstract500.

1.5.2 Evaluation Chapters

Chapter 9 (Evaluative) sets out a study design to evaluate the CVFM. It then describes a pilot study leading to amendments to the study design ready for the main evaluation.

Chapter 10 (Evaluative) describes the main study to evaluate the CVFM. In addition to reporting and reflecting on results of the main study, results from the pilot are integrated, take on further significance, and strengthen one area of the conclusions.

Chapter 11 (Evaluative), Discussion and Conclusion, discusses the implications arising out of the CVFM, the development of this particular implementation, and the results of the evaluations. The chapter ends with a summary of the thesis and ultimate conclusions.

Chapter 2

Motivation and Medium

There are several aspects to the motivation behind the development of the new mode of communication between designers and crowds (the CVFM). These include a) the observation that text dominates the conventional channels by which a designer can gain feedback from a network or crowd (with implications for the exclusion of some people due to factors of cognition) and b) another observation, that there is a well-established trend towards more participation by non-designers in the design process which increasingly requires designers to communicate beyond their colleagues. To explore these motivations, this chapter discusses several topics.

We start in Section 2.1 by describing some particular drawbacks of conventional feedback methods and how these can be avoided by the CVFM. Section 2.2 details potential drawbacks of the CVFM. Section 2.3 focuses on participatory design and how the CVFM can enhance this. In Sections 2.4 and 2.5 the usefulness of crowds, the challenge of summarisation posed by their use, and current provision of crowdsourced design feedback are discussed. Sections 2.6 and 2.7 examine the areas of cognition, emotion and imagery in design and how they relate to the CVFM. The last topic explored (in Section 2.8) is communication and semiology. The chapter concludes by summarising the outputs from these topic discussions.

2.1 Drawbacks of Conventional Methods of Feedback

In this section, the disadvantages suffered by conventional computer mediated methods of obtaining feedback are discussed. The conventional methods discussed here are surveys and feedback (or review) forums. How the CVFM can avoid or suffer less from the drawbacks suffered by these conventional methods is then discussed. Lastly, the drawbacks, and the position of the CVFM relative to them, are summarised in the conclusion to this section.

2.1.1 Surveys

Surveys are subject to biases. Causes of these biases include where

a) a portion of the population does not take part in the survey thus absenting some demographic from the results; this is termed, selective non-response (Maclennan et al., 2012); and

b) participants give answers which do not truly represent their own opinions but instead are answers which they perceive as being closer to social norms; this is termed social desirability response bias (Nederhof, 1985).

Selective non-response might occur in design feedback if a section of the population found text-based surveys demotivating and so did not take part. Social desirability response bias occurs due to two tendencies within individuals to answer survey questions insincerely (but not for malicious reasons). There are two aspects to this: "self-deception" and "other-deception" (Nederhof, 1985) where a survey respondent instead of answering truthfully gives an answer that they perceive as being closer to one fitting expectations or social norms. For example, "other-deception" might occur in design feedback if a respondent, concerned about hurting the designer's feelings, were to moderate their criticism. Survey respondents can worry about confidentiality (Tourangeau, 2001); i.e. that responses which are properly confidential might become attributed openly to them. This can lead to both of the previously described biases being accentuated.

2.1.2 Feedback Forums

Feedback forums also have some drawbacks: The picture they provide can be skewed by overly negative responses (Tuzovic, 2004). In addition they can contain polarised views and lack representation of moderate opinion. Contributors, by politicking, often try to have their view prevail in a forum (Talwar et al., 2007) and if this occurs the wisdom of the crowd (Surowiecki, 2004) would be compromised as it depends on the intellectual independence of each crowd member. A further feature of such forums is that online reviewers are discouraged from expressing emotions (Lee et al., 2008) as

subjectivity lessens the clarity of their message. This is despite the fact that the emotional impact of any product or design would be an important aspect of its success.

2.1.3 How the CVFM May Mitigate/Avoid These Drawbacks

The CVFM would suffer less (or differentially) from the drawbacks of surveys. Any selective non-response profile is likely to be different because potential respondents wary of conventional text surveys may find responding via images more appealing and so take part. Social desirability response bias could be less likely as, depending on the images used for feedback, a respondent might feel less accountable as image responses are likely to be open to interpretation. Confidentiality would be less of a concern if the content of a response consisted of images chosen from provided image banks.

Compared to feedback forums, the ability of feedback givers to give negative feedback via the CVFM would depend on its image repertoire. However, there would be no opportunity for politicking. Each contribution would have the same weight, in the same way as do the star ratings element of some review forums (Tsytsarau & Palpanas, 2012), thus contributors cannot argue for their point of view to become the prevailing one. Due to its visual nature spontaneity and subjectivity would be recognized as inherent in the medium of the CVFM, encouraging rather than discouraging emotion expression.

2.1.4 Conclusion to Section 2.1

The conventional methods of computer mediated feedback examined in this section (focussing on their drawbacks) were surveys and feedback forums. Two biases suffered by surveys, selective non-response and social desirability response were discussed. It was argued that the CVFM would suffer less from social desirability response bias due to the ambiguity inherent in images placing less of a burden of accountability on respondents; while any non-response profile might differ from or complement that of text-based formats due to the use of images attracting a different demographic. Two drawbacks of feedback forums, discouragement of subjectivity and the politicking by proponents of certain views, could not affect the CVFM as it would be recognised as inherently subjective and, like star ratings, would give equal weight to all responses.

2.2 Potential drawbacks of the CVFM

There will be drawbacks to using images for feedback. These are foreseen to be fall into two categories a) ambiguity of images, and b) those arising from downsides to crowdsourcing. The issue of ambiguity in images is addressed in Section 2.2.1 below. The issue of the downside to crowdsourcing is addressed in Section 2.4 Crowds and Crowdsourcing.

2.2.1 Ambiguity of images

There is often uncertainty over the semantic content of images. For example single images can evoke multiple emotions (Bradley, et al. 2001).

2.3 Co-Design / Participatory Design

In this section, in 2.3.1 and 2.3.2, aspects of the trend towards more involvement of users and customers in the design of products and services are discussed. In 2.3.3 how the CVFM might work as a tool to encourage participation in Virtual Customer Communities and the benefits that can bring are discussed. Finally 2.3.4 describes how the CVFM might allow the recording of co-design conversation and the value this would have.

2.3.1 "Prosumerism"

In 1980 Toffler, while highlighting new trends and making predictions of the way in which civilisation may be moving, coined the term "prosumer" (Toffler, 1980). A prosumer is a consumer who has taken on some responsibilities of work that was previously done for them in the production of a product or service which they simply consumed. This might be in an active mode such as the self-treatment of ailments or diagnosis at home using medical test kits previously only available to doctors. Or it might be something as simple as using self-service petrol pumps.

Co-design fits well with Toffler's idea of a prosumer. In co-design the consumer or user takes on some of the responsibility for the design (which is part of the process of bringing a product to market).

Many of Toffler's predictions (made from the trends he had observed up to 1980, pre-Internet) have come true. One in particularly is pertinent to this thesis; i.e. "*De-massification of the Media*", meaning people's interests are fragmenting into groups and their interests are becoming more specialised. This more individualised consumption of media is an aspect of more recent trends which include a move away from the consumption of print journalism to more diverse online outlets including blogging (Davis, 2009). It is hoped that the crowd in the CVFM will consist of such interest groups who may become aligned with specific designers.

2.3.2 Mass Customisation

An aspect of the idea of co-design is mass customisation (Piller et al., 2005). Some of today's prosumers are catered for by businesses which offer services allowing individuals to specify their own individual requirements. These requirements are usually limited to a set of options but the levels of customisation can be quite sophisticated. Adidas and Lego are examples of this (Piller et al., 2005). Adidas allows individualised specification of sports shoe construction and Lego allow a user to create their own Lego set with instructions and box graphics. However, this level of specificity in the customer's involvement contrasts with what is expected from the CVFM. The CVFM is intended to enable and engender a community or a following for designers who take on a leading (and more traditional) role in creating the design or prototype which they then develop informed by the crowds visual response.

2.3.3 Virtual Customer Communities

Virtual customer communities (online networks of consumers) or *VCCs* offer companies which cultivate them particular advantages if they engage customers closely in the design process and perhaps in co-design activities (Romero & Molina, 2011) (Porter et al., 2013) (Sanders & Simons, 2009). VCCs have been found to provide information on consumer behaviour and desires, resulting in savings on the research and development required to produce new products. They also generate increased brand loyalty.

The CVFM could be used as one tool to engage consumers in such VCCs based on low-effort cycles of co-design with crowds reacting visually to prototypes and improvements

while simultaneously growing loyalty to the designer or brand. Some of the crowd thus engaged might become available for more specific and literal market research.

2.3.4 Participatory Design Records

Sanders & Westerlund (2011) in an analysis of participatory design activities in physical (rather than virtual) settings identified the factors in co-design affecting its success. These included the design space (or environment), the experience and practice of those taking part and the consideration given to how the co-design activity is be recorded and communicated later. In the latter factor emphasising that effective co-design results in a record that can be shared among the interlocutors and disseminated. The CVFM, visual conversation cycles, taking place in a computer-mediated space will be easily recorded. It is envisaged that the record of each co-design process in our proposed system will add value to an associated final product in the form of an attractive visual narrative. Such added value does not involve consumption of additional physical resources in the way a garment or other consumer goods do (a significant problem in the fashion industry) and such value enhancements lead to environmental benefits through reduced consumption (Sanders & Simons, 2009).

2.4 Crowds and Crowdsourcing

This section focuses on the crowd aspect of this thesis. It begins by arguing that it is already recognised that there is value in the collective judgement of a crowd, notes that using the crowd necessitates summarising the crowd output, and clarifies that this thesis is indeed exploring crowdsourcing by defining the term.

2.4.1 The Judgement of Crowds

The idea that the judgement of the many might be superior to the judgement of a few experts is not a new one. Aristotle expressed it around 350 BC:

> *"For it is possible that the many, no one of whom taken singly is a good man, may yet taken all together be better than the few, not individually but collectively, ...For where there are many people, each has some share of goodness and*

intelligence, and when these are brought together, they become as it were one multiple man …" (Politics, Bk. III, Ch. 11, para. 1). Aristotle (1962)

He goes on to state that in this way the "many" are a better judge of works of music and poetry than the few.

A more recent and well known exposition of the idea is Surowiecki's *The Wisdom of Crowds* (2004), which gives many examples to support his thesis that the collective judgment of a crowd is often superior to that of an expert. He does make the provisos that for the crowd to be relied upon it must be diverse, large, and independent. The independence of each member of the crowd is a property to which Surowiecki assigns particular importance. Examples of when crowds have been seen to fail occur when independence breaks down and individual opinions are influenced by a group mentality such as during a stock market bubble.

Indeed in this regard one of the hoped for benefits of the CVFM, i.e. a designer building a following within the crowd (2.3.3), may cause a tension. A designer may value the "wise" judgement of the crowd and be expecting added value to accrue in any new design influenced by crowd feedback. However, if a designer is successful in building a following, and that following begins to form a group mentality, then the "wisdom" of the crowd could be compromised. If this does happen it may be that the value added (and potential purchases) within the designer's following will outweigh any loss of global value in the finished design due to any degradation in the global crowd's judgement. However, it may be sensible for the feedback from the designer's following to be collected and analysed separately from the global crowd. The designer could receive two separate streams of feedback and make design decisions accordingly.

Thus we see that there is value to be gained from the opinions of a crowd but that the independence of the individuals that constitute the crowd should be guarded to protect that value and to prevent the crowd becoming a liability rather than an asset.

2.4.2 Aggregation and Summarisation

Another feature of the wisdom of the crowd is that for it to be accessible the judgement of the crowd must be able to be aggregated or coalesced in some way such that it can be interpreted and acted upon. E.g. Surowiecki's (2004) opening example is one he draws from Galton's 1907 article in *Nature* "Vox Populi (The Wisdom of Crowds)" (Galton,

1907a). Galton, although at the time personally sceptical about crowd wisdom, showed that the median value estimated for the "dressed weight of an ox" taken from the 787 estimates submitted by a cattle show competition crowd was less than one per cent adrift from the true value. Galton in another article in the same issue of Nature proposed the use of the median value (from 12 suggested values) if a jury were deciding on a figure for damages in a court case (Galton, 1907b). Another example of useful crowd wisdom, market prices, tends to be consumed as single price figures or as single figures with a trend (down or up) rather than as the multitude of recent deal prices over a given period up to the present.

Thus, for use to be made of the value in a crowd's collective judgement, the judgements of all the individual crowd members must be able to be aggregated or summarised such that the judgment can be consumed or read conveniently. For Galton, when the "crowd" were individually each contributing a number, it was the median that he advocated as the way to access what Surowiecki would call the "wisdom" within the crowd's data. For the CVFM, with each crowd member contributing images, the challenge will be to summarise the totality of the crowd's images into a meaningful but concise form.

2.4.3 Are We Crowdsourcing?

The comprehensive definition of crowdsourcing accepted by this thesis is quoted below:

> *"Crowdsourcing is a type of participative online activity in which an individual, an institution, a non-profit organization, or company proposes to a group of individuals of varying knowledge, heterogeneity, and number, via a flexible open call, the voluntary undertaking of a task. The undertaking of the task, of variable complexity and modularity, and in which the crowd should participate bringing their work, money, knowledge and/or experience, always entails mutual benefit. The user will receive the satisfaction of a given type of need, be it economic, social recognition, self-esteem, or the development of individual skills, while the crowdsourcer will obtain and utilize to their advantage what the user has brought to the venture, whose form will depend on the type of activity undertaken." (Estelles-Arolas and Gonzalez-Ladron-de-Guevara, 2012).*

Can we term what is set out in this thesis "crowdsourcing"? There are two aspects of the thesis which are suggested to be crowdsourcing:

a) The design feedback in the CVFM is eventually intended to be sought from an Internet crowd.

b) Some of the judgments used to build the image browsers developed in this thesis (Chapter 4 and Chapter 8) are gathered through crowdsourcing services.

By the quoted definition, in b), above, our use of crowdsourcing services is definitely crowdsourcing. Likewise so is a), but with the proviso that the recruitment of the crowd be of a sufficiently open nature; i.e. not restricted to a particular group.

Thus, according to the quoted definition of crowdsourcing, the CVFM and the sourcing of the perceptual data used in this thesis can be validly termed crowdsourcing.

2.4.4 The Ethics of Crowdsourcing

There is a downside to crowdsourcing in that it has allowed a degree of what is seen by many as exploitation of crowdsourced workers by low paying crowdsourcing platforms such as Amazon Mechanical Turk (AMT). Often the pay for workers is well below the UK and USA national minimum wage levels. (Schmidt, 2013) (Horton & Chilton, 2010) (Brabham, 2012). Indeed there are two main aspects to the exploitative practices on platforms such as AMT. One is the low pay for micro-tasks or HITS (Human intelligence tasks). The other is the practice of having workers such as graphic designers compete against each other. Work providers, clients of the crowd platforms, commission work such as a graphic design brief; workers (designers) operating as freelancers do the work, effectively gambling on getting paid, and then submit it competitively with only one piece of work being accepted. Workers who lose the competition do not get any pay (Schmidt 2013). Such practices are routine on crowdsourcing platforms. In this thesis crowdsourced workers are engaged for the image sorting and categorising tasks. Due consideration is given to calibrating the pay to be commensurate with expected time on task and the UK national minimum wage.

There is, however, another and positive side to the ethics of crowdsourcing. That is through the engagement of volunteer crowds where the individual members are motivated by altruism and community spirit. One example of this is a collection of projects related to astronomy research called *Zooniverse* including *Galaxy Zoo* which has participants identify astronomic objects and phenomena in images. (Savage, 2012) Another is *Wikipaedia* the online encyclopaedia. Although these projects make use of volunteers there does have to be a benefit for those in the crowd who work on them

(Savage, 2012). These benefits can be purely altruistic such as in Wikipeadia's case or there can be an actual payback for participants such as learning a language while this activity serves a translation purpose as in Von Ahn's *Duolingo* project (Garcia, 2013).

In the case of the CVFM it is hoped to engage a crowd through altruism and social involvement as the motives. Thus the CVFM is not to be part of the exploitative side of crowdsourcing. Even the eventual continuation of image categorisation beyond the research investigation work of this thesis into an actual live implementation is expected to engage participants in image categorisation for fun as reward rather than monetary pay. This can be through gamification (von Ahn & Dhabish 2008) or community spirit as in Dribble (Cook et al., 2009) or Wikipeadia.

2.4.5 Conclusion to Section 2.4

This Section has argued a) that it is recognised that there is value in the collective view of a crowd and that it can be as good or even better than consulting an expert b) that for the "wisdom" of the crowd to be useful and accessible, the individual views of the crowd members must be able to be aggregated or summarised such that the crowd view can be consumed and acted upon*, thus identifying one challenge for this thesis, that of summarising the image selections of a crowd*, and c) that what this thesis proposes can indeed be validly termed "crowdsourcing". Additionally, in Section 2.4.4, the ethical pros and cons of crowdsourcing were discussed, exposing the exploitation that goes on via crowdourcing platforms and the altruistic volunteering which conversely does take place in projects such as Galaxy Zoo and Wikipeadia. That section concluded by noting that the CVFM was not intended when implemented live to use crowdsourcing platforms such as Amazon Mechanical Turk but instead follow the volunteer or gamified model as exemplified by Wikipeadia and von Ahn & Dhabish's (2004) *ESP* project.

2.5 Crowdsourced Design Feedback

In this section, current provision of crowdsourced feedback specifically for design is discussed. First blogging and feedback forum communities are discussed. Then more recent specific crowdsourcing tools are described.

Blogging or involvement in communities such as *Dribbble* (2015) and *Reddit* (2015) has given designers access to feedback from crowds. However, the level of commitment required to participate in such online communities (Cook et al., 2009) limits their accessibility. These methods can suffer from the drawbacks described in 2.1.

Specific tools have been created for crowdsourcing feedback (Xu et al., 2014) (Luther et al., 2014). These allow paid participants to be engaged by designers on services such as CrowdFlower (2015). For example *Voyant* (Xu et al., 2014) is a crowdsourcing tool for efficiently obtaining, specific, objective, feedback on graphic designs from paid crowdsourced workers in a structured way avoiding the need for a designer to have expertise in constructing the human intelligence tasks required by services such as CrowdFlower. Xu et al., (2015) have gone on to show, through a linguistic analysis, that the structured feedback from crowdsourced workers used significantly less emotion words compared to free form feedback. They concluded the structured feedback was significantly more deliberate i.e. less spontaneous.

The CVFM is intended to complement rather than compete with such systems by encouraging the participation of volunteer crowds, perhaps engaged through social media, and seeking subjective mood-style feedback.

2.6 Cognition

Two aspects of cognition are considered here. First, cognitive styles (how individuals tend to process and internally represent information) are discussed and related to the intended medium of the CVFM (i.e. images). Intuition, its importance, and the theoretical prospect of it being exploited by the CVFM, are discussed. The difference between cognitive styles and intuition is noted along with possible cultural influences on cognitive style. Emotion is introduced as an important factor in cognition. Lastly, the conclusions from these discussions are summarised.

2.6.1 Cognitive Styles

Cognitive styles have been used to inform teaching and learning (Coffield et al., 2004) and are often used to predict people's performance in different circumstances (Kozhevnikov, 2007). Research in the field of cognitive styles produced several models (Rayner & Riding, 1997). However, two main cognitive style dimensions were

identified by Riding & Cheema (1991) in a review: "wholist-analytic" and "verbalizer-imager" (Figure 2.1). The "wholist-analytic" dimension describes whether an individual tends to process information in wholes or in parts. However, more pertinent to the CVFM, is the verbaliser-imager dimension which categorises people as either tending to represent information during thinking verbally or in images (i.e. they lie on a continuum between these two modes).

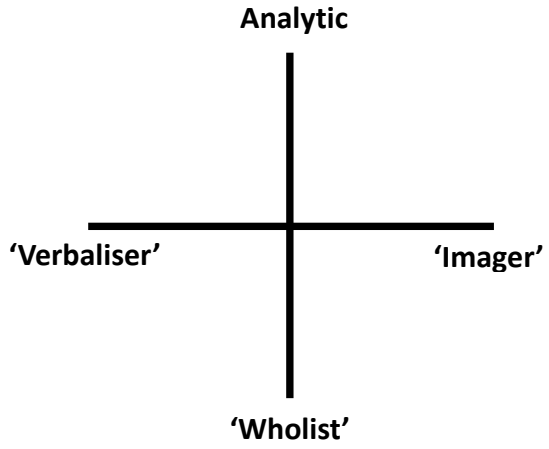


Figure 2.1- The two main cognitive style dimensions. (Adapted from Riding (1997)).

While cognitive styles are independent of gender, intelligence, and age (Riding, 1997), differences between visual and verbal people have been measured in brain activity patterns. These have been observed to be different in visual and verbal individuals when engaged in certain tasks (Gevins & Smith, 2000). In terms of information consumption, visual people learn better when consuming information pictorially rather than verbally (and textually) (Riding & Ashmore, 1980).

It is the work in this field that lies behind a broad acceptance that some people prefer, and are more suited to, consuming information visually rather than verbally (or textually).

The CVFM, being based on responding with images, is expected to appeal especially to crowd users who are of a visual cognitive style (on the "verbalizer-imager" dimension) thus providing a channel which has the potential to attract more people into design feedback than would be the case were design feedback to continue to depend on text-based methods. It is possible that crowd members who are more visual than verbal can provide feedback which designers would find more valuable. (Indeed this is discussed in 10.5.2).

2.6.2 Intuition

Another aspect of cognition, intuition, was characterized by Plato (circa 380 BC) as being the highest form of thought (Plato, 1998). Intuition has since been defined in work on dual process theory.

Dual processing theories (Epstein, 1994) (Sloman, 1996) (Evans, 2003) are used to explain the apparent duality in the way that people make decisions; some decisions are arrived at slowly following a logical and analytic process, whereas intuition leads to a fast, almost effortless conclusion. The theory terms the fast intuitive process as System 1 and the slow deliberative process as System 2. A rationale for the existence of these parallel systems is that we have recently evolved the specifically human System 2 while still possessing System 1 from our more distant evolutionary past (Evans, 2013). Experts are thought to often apply System 1 as it exploits previous experience (Evans, 2008). It might be expected that the slow analytical System 2 process would lead to more accurate outcomes but the fast, System 1 can equal and sometimes better System 2 in terms of the quality of the answers produced (Norman et al., 2014, Witteman et al., 2009).

Evans (2008) describes system 1 as automatic, low effort, rapid, holistic, perceptual, *nonverbal* and independent of working memory. Whereas he characterises system 2 as controlled, high effort, slow, analytic, reflective, *linked to language*, and limited by working memory capacity.

It is interesting to note that the dual process dichotomy came to be more fully recognised when analytical thinking in decision making was shown to be plagued by biases due to the unconscious encroachment of intuition on what, the individuals concerned believed, were entirely logical thought processes (Evans, 2003). It is actually System 1 that takes care of most of our every-day decision making (Evans, 2013).

These findings show that there is a pervasive and embedded nature to System 1 and intuition. Taking account of this and embracing intuition is to embrace human nature. The purpose of the CVFM is to encouraging intuitive, perceptual and nonverbal feedback. The next section discusses how certain we can be about being able to engage intuition using the CVFM.

2.6.3 Cognition: Styles, Types and Culture

We cannot conflate here cognitive styles and dual processing theories. These two aspects of cognition are deemed separate (Evans and Stanovitch, 2013). System 1 and System 2 are *types* of cognition and are considered to be more deeply rooted than cognitive styles which a) are considered to be different styles of System 2 *type* thought processes and b) may have environmental in addition to physiological roots. Holistic-analytic styles have been observed to differ between cultures. Nisbett et al. (2001) compared holistic-analytic cognition across cultures and found that individuals from East Asian cultures tend to be more holistic in cognitive style whereas those from western cultures tend to be more analytic in style.

Thus there are two points here of significance for the CVFM. Firstly, while imagery may help prompt the use of intuition due to the non-verbal nature of System 1, it may not be possible to do anything other than speculate about whether or not intuition is in fact engaged during use of the CVFM. On the other hand the verbaliser-imager cognitive style dimension does appear to be a factor which is likely to affect the appeal of the CVFM among potential users. Secondly, images as a medium are clearly going to be language independent; thus if we are theorising that the CVFM will have varying appeal depending on cognitive style in users, then that might also equate to a varying appeal depending on the cultural background of users. (Although it is just the "wholist-analytic" dimension, and not the "verbaliser-imager" dimension, that has been demonstrated to vary with culture.)

2.6.4 Emotion in Cognition

Emotions are recognized in the literature as playing a role in intuitive thinking, decision making, and information processing (Schwarz et al., 1991) (Tiedens et al., 2001) (Lerner et al., 2004). For example it has been shown that exposing subjects to different emotion stimuli prior to eliciting buying and selling decisions from them, had a dramatic effect on the values placed on items bought and sold (Lerner et al., 2004). Indeed Mikels et al. (2011) showed that for complex decisions using a feelings-based approach produces better quality decisions.

Thus if the CVFM can be used to access peoples emotional reaction to a design, given that emotions can affect buying decisions the CVFM may have impact as a marketing tool.

2.6.5 Conclusion to Section 2.6

In *styles* of thinking: of the two main cognitive style dimensions' one determines whether individuals prefer and are more suited to consuming information visually rather than verbally (or textually). Cultural influences may influence cognitive style.

In *types* of thinking: intuition is recognised in Dual Process theories (System 1 and System 2). System 1, *fast*, *intuitive*, *non-verbal*, can, arguably, be the more productive of the two systems, indeed it also influences logical System 2 thinking.

Emotion is recognised to affect cognition; in particular affecting decision making such as purchasing decisions.

The CVFM can appeal to potential users of a visual cognitive style by offering images as its medium thus better catering for users of this style compared to more verbal users who are already well catered for by conventional methods. It might also be possible to encourage use of intuition in feedback by use of images, in that discouraging use of language (linked to System 2) might prompt users to resort to intuition (System 1). If possible, the mode of image selection should encourage the deployment of intuition. Consideration should be given to the communication of emotions using the CVFM.

2.7 Emotion, Mood Boards and Imagery in Design

2.7.1 Emotion in Marketing and Design

As mentioned already in 2.6.4 emotions have a role in intuitive thinking, decision making, and information processing. It is perhaps not surprising then that designers are interested in emotions. Approaches such as Kensai engineering (Nagamachi, 1995) directly take emotions into account in the design process.

Lim et al. (2008) examined emotion and product design. They categorised users' emotional responses to products as falling into three categories: Visceral, behavioural, and reflective. The visceral are based on perceptions (appearance), the behavioural are

based on expectations e.g. frustration when these are not met, and the reflective are associated with experience and might involve reactions to emotions experienced during visceral and behavioural phases of emotion response to a product.

As the CVFM is expected to be used in a prototype development mode it is less likely that the behavioural phase emotions will be involved. However, the CVFM could provide access to the visceral phase emotions, and might tap reflective phase emotions if a design were a progression from an earlier one.

Emotions are also recognised as important in marketing (Taylor, 2000) (Mizerski & White, 1986) with their influence on decision making already having been mentioned in 2.6.4.

2.7.2 Mood Boards, Images and Emotion

The importance of images in establishing and developing a perceptual and emotional theme (or mood) for a design is recognized in the design practice of mood boards. Indeed mood boards (see Figure 2.2) are a conventional way in which designers gain inspiration. They are a well-established creative and analytical tool used by designers when creating a design idea (Eckert & Stacey, 2000). With Mood boards, designers use images and objects to develop a perceptual and emotional theme. The images can be chosen purely for their visual properties but they can also be included because of their cultural content where a cultural feature is an aspect of the design theme. Although a design mood can be described in text, such a description is inherently sequential, whereas the mood portrayed in a mood board can be engendered as a whole simultaneously and in one view. Also, to avoid specific figurative connections, abstract images are often used (Garner, S. & McDonagh-Philp, 2001).

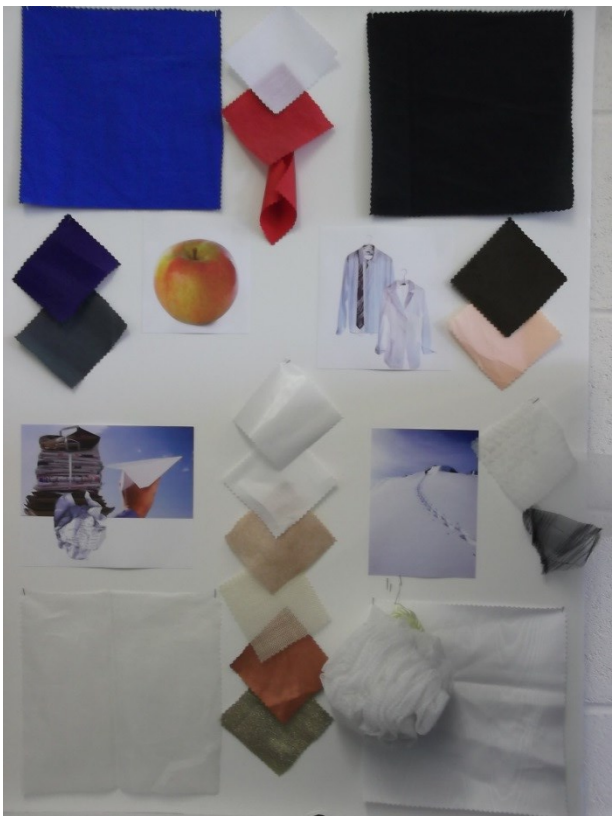


Figure 2.2 – Example of a mood board

However, figurative images can access emotions in a more specific way than can abstract images (Bradley et al., 2001). Mikels et al. (2005) categorized images according to their emotional affect. There is a good prospect of emotive images being suitable for fast intuitive feedback because it has been shown that people rapidly and reliably interpret the emotion content of images (Junghöfer et al., 2001).

Thus both abstract and emotive imagery may usefully be considered for use in the CVFM.

2.7.3 Conclusion to Section 2.7

The emotional impact of designs is accepted as important as illustrated by design practices including the use of mood boards. Emotions are also important in marketing. The common use of abstract images in establishing a "mood" for a design is practical recognition of the utility of images in conveying emotion. Both abstract and emotive imagery may be appropriate for use in the CVFM.

2.8 Communication and Semiology

This section examines aspects of communication. First those aspects not directly concerned with the intended meaning in communication are discussed. Then whether or not meaning will be able to be conveyed successfully in images or pictures is addressed. Lastly conclusions about these issues with respect to the CVFM are summarised.

2.8.1 Communication

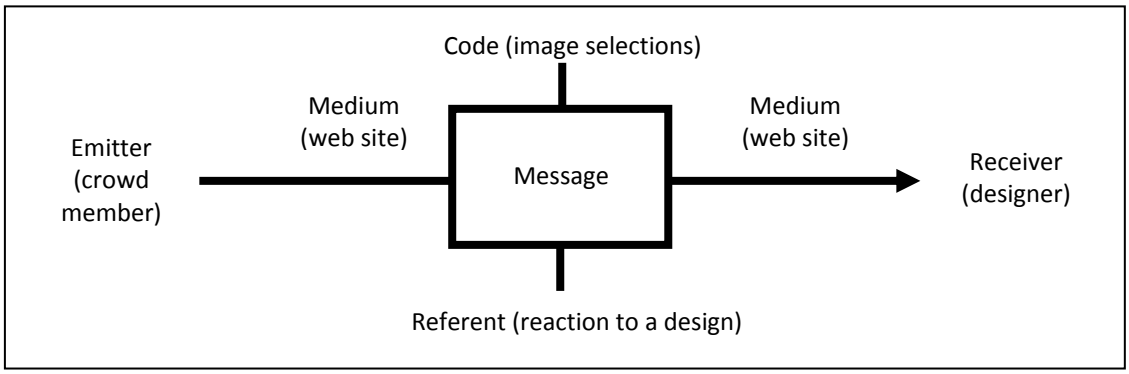


Figure 2.3 - Diagram of the main actors and elements of communication in the CVFM (Adapted from Guiraud (1971)). (Although no specific implementation has been decided for the CVFM we suggest here that it may be done using a web site.)

Figure 2.3 sets communication issues in the context of the CVFM for discussion. There are two aspects to communication which will be addressed. First, perhaps the obvious aspect is the meaning in the communication: i.e. what is overtly said and read by the parties (designer and crowd). The designer will be showing their design and that will be in some form of image or presentation. However, nothing innovative is planned for the CVFM in this regard so we will be focussing on what the crowd will say to the designer, and in particular, how the crowd will say it and how the designer will read it, this being the innovative side of the CVFM. Secondly, and perhaps, less obvious is the process of communication itself and aspects intrinsic to communication which may come to the fore in the CVFM.

The issue of how the crowd will use images to describe its reaction to a design and how the designer will read that, i.e. issues of semantics and semiology, will be addressed later in 2.8.2.

However, addressing now some generalities of communication, the semantics of a message is not the only reason for a conversation. Jakobson (1960) set out aspects, or as

he labelled them "functions", of communication. Three are pertinent to the CVFM and are set out in Table 2.1.

"Function"	Significance to the CVFM
A message can be aimed to illicit a logical or an emotional response from the recipient.	a) The designer hopes the design will provoke an emotional reaction; b) Perhaps some crowd members will be provocative in their image choices in reply?
A message can have its own intrinsic artistic or poetic meaning.	Some visual feedback summaries are likely even to possess their own intrinsic artistic meaning or merit and if so, designers receiving them may benefit from this in terms of inspiration either to change their design or in ideas for a further design.
An important and common function of communication is simply to continue the conversation; i.e. the semantic content can be entirely superfluous to its purpose.	The visual conversation would be a manifestation of a relationship between designer and crowd and as such have value in its own right.

Table 2.1 – Three of Jakobson's (1960) "functions" of communication and the significance of each to the CVFM.

These non-semiotic properties of the visual feedback conversations that the CVFM will enable may be as important as the purely semantic message content of image based crowd feedback.

2.8.2 Semiology

Chandler (2002) defines semiotics as

> *"…the study… of anything which 'stands for' something else."*

Aside from three aspects of communication described in the previous subsection, an important aspect of any conversation between a designer and a crowd will be whether or not the designer will be able to understand what the crowd has said in its images; i.e. (referring to Figure 2.3) will the crowd (the *emitter*) be able to successfully encode its reaction (the *referent*) in its image selections (the *code*) and will the designer (the *receiver*) be able to read it (or decode it)? Also will the visual summaries carry the same message? These questions are addressed experimentally in Chapter 7, however here some theory concerning this issue is briefly discussed.

Sausseure, in his theory of language, as described by Culler (1976), argued that, in language, signs are an arbitrary combination of signifier and signified; e.g. there is no natural reason for the word, dog, to signify what we recognise as the furry animal that

barks. Thus, if the images used for communicating are considered to be a totally new language, this raises the prospect of an involved and time-consuming language learning process to be gone through before the crowd and designers can communicate. However, it is expected that the image set(s) or visual medium will capitalise on current visual conventions already within the experience of the crowd and the designers and so allow communication both to take place initially and to develop.

Visual communication is already often done with pictographic symbols and icons. Signs without words at airports and on our roads are clear evidence that symbolic visual communication works. Neurath (1936) developed a language of pictures (or icons) to be used in education. Indeed pictographic languages such as Japanese use characters originally derived from stylised drawings. A communication channel using established signs or emoticons could be a valid component of a visual feedback system. However, the proximity of pictographic symbols and emoticons which are in common use already means that there would be less novelty and challenge in using these as a medium for the CVFM.

Hebecker & Ebbert (2010) have investigated the development and recognisability of free-drawn symbols in response to stimuli terms using a Pictionary-like online game. Free-drawn sketching as a medium for communication during the co-design process is recognised as valuable (Craft & Cairns, 2006), so the possibility of including a free-drawn sketch application such as that used by Hebecker & Ebbert (2010) as a channel for design feedback could be used for the CVFM. However, hand-drawn sketches, although useful as a medium for collaborative design, would present two difficulties for the CVFM a) summarisation of aggregated sketches would pose a challenge and b) the moderation of crowd responses would require consideration. There will always be a mischievous element in any online society whose input may offend other users (Kirman et al., 2012). Moderation of responses is a problem in text-based systems where users are free to type their own possibly offensive words. Offensive hand-drawn sketches would be more difficult to moderate than text (which can have some automated filtering.)

2.8.3 Conclusion to Section 2.8

From the “functions” of communication as set out by Jakobson (1960), it is expected that the CVFM may 1) enable designers to provoke the crowds emotional reaction to

design; and perhaps allow the crowd to provoke some design changes by the designers, 2) lead to visual summaries possessing intrinsic artistic meaning or merit and inspire designers to make changes or produce new ideas; and 3) enable a relationship between designer and crowd.

Any sign can be used to signify any signified thing and so there will be scope for new language to develop. However it is expected that the medium used for the CVFM will capitalise on existing visual conventions to establish initial communication.

Hand drawn sketching and emoticons were considered for the CVFM but the problem of offensive or unsuitable input from the crowd and lack of novelty counts against these.

2.9 Conclusion to Chapter 2

The conclusions from the sections of the chapter are summarised below:

Compared to surveys the CVFM is expected to suffer less from social desirability response bias and have at least a different and perhaps even reduced non-response profile due to the use of a visual medium. The equal weight of each visual response in the CVFM would avoid the politicking and polarisation of views which can occur in feedback forums.

The well-established trend of consumers becoming more involved in the creation of products ("prosumerism") now manifests itself in virtual customer communities. These have benefits for businesses and designers and could be enhanced by using the CVFM as a tool to encourage additional participation. The recording of the CVFM's visual conversations could add value to any end products with attractive visual narratives. Indeed the idea that there is value (or "wisdom") to be had from the output of a crowd of non-experts is recognised, but raises the challenge of summarising the crowd's visual output in the CVFM. The CVFM is expected to complement existing crowdsourcing tools for design feedback rather than compete with them.

Visual cognitive style and intuition can be exploited by the CVFM. Emotion is important in design and marketing; imagery is already used in design to access emotion. Pictographic symbols especially emoticons are already in common use but offer less novelty as a medium. User generated sketches would pose a problem of detecting unsuitable crowd input. However, abstract and emotive imagery would be suitable

candidates for use in the CVFM and these should be deployed in a way that encourages use of intuition.

Lastly, while the use of images to form messages poses a semiotic challenge, it also offers the prospect of inspiration for designers from intrinsic artistic value in visual feedback generated via the CVFM.

Together these topics show there is both a place for, and a gap in, the provision of channels connecting designers with crowds in feedback. Indeed that gap amounts to an asymmetry in design communication. Much of what designers express about their designs is visual yet most feedback that designers would currently expect to receive from networks or crowds is textual. There is a lack of an engaging, visual and yet practical medium for communication between a crowd and an individual designer. The crowdsourced visual feedback method developed in this thesis can step into that gap using imagery as its medium.

Chapter 3

Interface for Image Selection

Given that the previous chapter concluded that images are to be the medium for the CVFM, crowd users will require an interface via which to select images to form their visual responses. The purpose of this chapter is to establish the appropriate format for that image selection interface.

Table 2.1 sets out the requirements for the image selection interface along with their motivations.

ISIR No	Image Selection Interface Requirements (ISIR)	Motivation
1	Intuitive image selection	One conclusion from Chapter 2 was that the CVFM should encourage use of intuition.
2	We must own the images or be allowed to use them	Should an image, which was the subject of copyright, appear in the feedback without the owner's permission we could face being invoiced for its use. We want our implementation of the CVFM to be able to be openly accessible.
3	Use a closed set of images	If the image pool from which the feedback is drawn were outside the system's control the probability of unsuitable images entering the feedback would exist. This could be damaging and should be avoided. (cf. 2.8.2).

Table 3.1 - Requirements for an image selection interface with motivation for each.

Section 3.1 introduces content based image retrieval and the disadvantage of query-based search in the case of the CVFM. Section 3.2 examines browsing concluding that it will be the better approach for this thesis. Section 3.3 addresses the issue of how to structure any image set used by the CVFM to enable intuitive browsing. In addition the extraction of computer vision features is discussed and the problem of the "semantic gap" is exposed and described. Lastly, section 3.3 concludes that the semantic gap problem can be avoided by using human perceptual data. Section 3.4 describes various methods of gathering perceptual data all of which face limitations on the size of the image set with which they can be used. Then in Section 3.5 a method of obtaining perceptual data on large image databases is described along with an intuitive browsing environment which can a) exploit such data and b) has been shown to be superior to two

other browsers using this type of perceptual similarity-based browsing. Finally in the conclusion, requirements for the image selection interface are revisited to show that the chosen browsing environment satisfies them.

3.1 Content Based Image Retrieval (CBIR)

The problem of users needing to locate images within a large database of images has led to the study of content based image retrieval (CBIR). From a user's point of view there are two basic approaches that are used: query by example (where the user provides an example image as a query) and relevance feedback (where the user, over several interactions, narrows down the system's "search" results by describing each as relevant or irrelevant). However, both these approaches require the user to have a fixed definite query image either to hand or in mind. The next section examines browsing as an approach which addresses issues related to this.

3.2 Browsing

Heesch (2008) pointed out several advantages of browsing over query-based image search. Those particularly relevant to the CVFM, given its requirement for intuitive image selection (ISIR 1, Table 3.1) are summarised in Table 3.2.

Category of advantage	Details
Fluid information needs	As mentioned in 3.1, users may not have a particular image in mind. An initially vague requirement can develop and clarify during interaction with the database.
Mental query	In CBIR if a query image is required then this necessitates having images to hand for likely queries. Some systems allowed users to sketch a query, but this a) requires special input devices, b) skill in their use, and c) graphic expressive ability. Prior tagging of images with words which can then be used in a query. However not all visuals can be adequately described in words. Browsing can allow a mental query to be satisfied.
Exploiting the cognitive abilities of users	The human visual system can recognise patterns quickly and reliably. Browsing systems can harness this cognitive ability by facilitating fast decisions on relevance by users.

Table 3.2 - Advantages of browsing over query-based image search (Heesch, 2008).

The advantages of using a browsing strategy (rather than a query-based search) for the CVFM are clear. Thus the interface for image selection in the CVFM will be some form

of browser. It remains, however, to decide on the method of structuring any image database to enable intuitive browsing. The next section examines that issue.

3.3 Structuring an Image Set to Facilitate Browsing

Heesch (2008) points out that in any collection we expect the collected objects that are similar to each other to be near to each other and accessible from one another. Thus browsing systems depend on data which describes the similarity of images within their databases. Commonly this similarity data is gained by extracting features from the images using computer vision algorithms. Section 3.3.1 briefly discusses computer vision features and introduces the idea of the "semantic gap" which is defined in section 3.3.2. Section 3.3.3 points out work which has acknowledged the inadvisability of relying on computer vision features for image similarity data by contrasting such data with human perceptual similarity data. Lastly, 3.3.4 concludes with the decision that human perceptual data should be the basis of the image browsing for the CVFM.

3.3.1 Computer Vision Features

Computer classification of images for CBIR is based on feature extraction done by analysing the image content in terms of colour, shapes, edges, regions, objects etc.

Colour is one of the simpler features to process and this can be enhanced by processing luminance along with it (Keriminen & Gabbouj, 1999, 2000) (Keriminen et al., 2000). Chen et al. (2000) discussed the features used in content based classification of images and the table below briefly summarises that discussion.

Feature	Source of feature data.	Pros and Cons
Colour	An image's colour histogram, i.e. the frequency of pixels of certain colour bands in the image.	Low storage and simple computation requirements. Not affected by rotating the image.
Texture	Varying methods used including texture spectrum (He & Wang 1990)	More spatial detail than from colour histogram analysis.
Shape	Varying methods including edge orientation and distance transform.	Computationally expensive. Only low level shape features can be reliably extracted. This can be used to combine special detail with colour histogram analysis.

Table 3.3 - Summary of the discussion of features used in CBIR (Chen et al., 2000).

However, whatever the relative merits of these various features, matches based on these often do not bear a sematic resemblance in terms of topic (Sharma & Singh, 2011). This problem in computer vision has been identified as the "semantic gap" and is described further in the next section.

3.3.2 The Semantic Gap

Smeulders et al (2000) pointed out that there is a problem with relying on computer vision features to provide similarity data due to the "semantic gap" between what can be extracted from an image's features compared to what that image actually means to a user when viewing it. In short, automated computer vision does not match human perception in all its semantic complexity.

3.3.3 Computers vs. Humans in Judging Image Similarity

What the automated image processing methods used for CIBR are seeking to do is replicate human perception of the images being classified. The reservations about computer vision features expressed by Smeulders et al (2000) were presaged by Rogowitz et al. (1998). That work compared multi-dimensional scaling (MDS) visualisations of a set of photographic images based on human judgments of their similarity, with MDS visualisations from image processing algorithms. The results led to the conclusion that the automated image processing similarity metrics were not an adequate model of the perceptual data.[2] More recently, Depalov et al (2006) state that CIBR systems are still unable to match human perception. Also Clarke et al (2011) showed that for texture images, similarity data based on computer vision features did not match data from human judgements.

3.3.4 Conclusion to Section 3.3

Taking into account a) these reservations about relying on computer vision features, b) the requirement for intuitive image selection, and c) the prospect of being able to obtain perceptual data on any image set used for the CVFM due to the other requirement for a

[2] One interesting finding in the paper was that the MDS visualisations of the data suggested a "man-made vs. natural" axis within the view of the perceptual data. This was something clearly to be seen in the MDS visualisations of the *Abstract500* perceptual similarity data in this thesis (See 4.6).

closed set of images (ISIR No. 1 and 3, Table 3.1), we conclude that a browser for the CVFM should be organised based on human perceptual data. The next section discusses how such data might be obtained.

3.4 Methods of Gathering Perceptual Visual Similarity Data.

If the CVFM image browser is to use perceptual data then it is appropriate to examine methods for obtaining this type of data on an image set. Four methods are summarised in Table 3.4.

Method	Description	Pros and Cons
Table sorting (Rogowitz et al., 1998).	Observers place images on a table arranging them such that images most similar to each other are close and those dissimilar are far apart. The distance between each pair is measured.	Difficult to record and there will be some practical limit on the size of the image set. Other problems are a) both observer time and fatigue and b) table size.
Paired comparisons (Rogowitz et al., 1998).	Observers view a pair of images and assign a number proportional to the judged similarity.	The number of pairs grows with the square of the number of images in the set, rapidly becoming too large to contemplate one observer viewing them all. The subjectivity in the observer's similarity score can lead to bias. A modified version was used by Rogowitz et al. (1998) Which removed the subjective score element, reduced the number of observations needed and perhaps would be considered a version of the pairs of pairs method.
Pairs of pairs (Clarke et al., 2012)	Observers view two pairs of images and nominate one of the pairs as being more similar to each other that the other pair are.	Very time intensive. As above, the number of pairs grows with the square of the number of images in the set, rapidly becoming too large to contemplate one observer viewing them all.
free sorting (Clarke et al., 2012)	Observers group images on a table into piles of images which they deem similar.	The size of the image set that can be practically free-grouped by one observer is limited, but less time intensive than the pair comparison methods if the set size is limited.

Table 3.4 - Methods of gathering perceptual visual similarity data.

Table 3.4 shows that the size of the image set is a major factor seen as limiting all of the methods described. However, recent work on texture image browsing environments by Halley (2012) has addressed this issue and is described in the next section.

3.5 Scalable Large Image Database Browsing using Perceptual Similarity

Halley (2012) (also described in Padilla et al. (2013)), in seeking a solution to the problem of producing a browsing environment for texture images which does not suffer from the mismatch between computer vision and perception, developed a method of obtaining perceptual data on a large database of 500 images. The method uses standard lab-based free sorting (Table 3.4) for a subset (100) of the images which informs the construction of a browser termed the bootstrap browser. (See 4.5.5 and 4.5.6). The bootstrap browser is then used as a structure allowing further similarity judgments to be gathered by engaging hundreds of crowdsourced participants to each liken a small number of the remaining 400 query images to images to be found in the bootstrap browser. This process produced a 500x500 similarity matrix which described the similarity relationships between all 500 images in that database.

Having obtained this perceptual data Halley (2012) went on to use the 500x500 similarity matrix to inform the creation of three browsing environments including one designed by Rogowitz et al. (1998) already referred to in Table 3.4 and another by Wittenburg et al. (1998). Experiments in which participants were tasked with finding given query images using the browsers showed one of the tested browsers to be superior to the other two. That superior browser was one which uses a rectangular self-organising map format (Kohonen, 1990, 1998) (Vesanto et al., 1999).

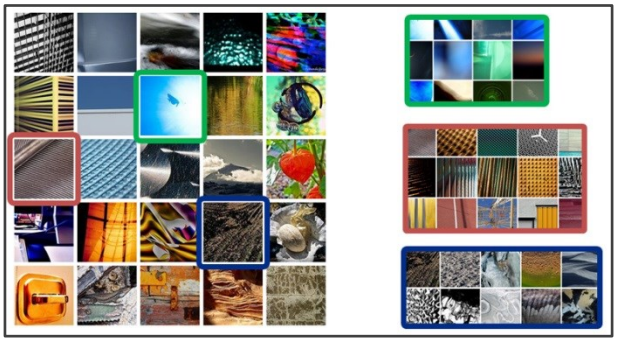


Figure 3.1 –A rectangular SOM browser presenting a large abstract image set in 5x5 stack configuration (left) with samples of images from three of the stacks (right).(The image set loaded in this example 5x5 stack browser is the abstract set gathered in Chapter 4).

What makes the rectangular SOM browser so successful is the way in which its layout is intuitive. (Figure 3.1 shows an example). It presents an array of image stacks in which each stack contains images which are highly similar. Stacks open when tapped or

clicked. Adjacent stacks hold images that are quite similar while stacks far apart on the array contain images that are dissimilar. Each stack represents a cluster of images (in terms of the perceptual similarity) with the top image being that nearest the cluster centroid by Euclidean distance and the rest of the stack listed in order of distance from the centroid. It is possible to deploy the rectangular SOM in configurations which vary the dimensions of the top level array of stacks, e.g. 5x5 (5 rows and 5 columns of image stacks) or 8x6 etc. The flexibility of the rectangular SOM browser was demonstrated in Padilla et al (2013) in which the rectangular SOM browser was shown to continue to offer its superior performance in a number of different stack configurations and with different screen sizes.

Thus in the next section, the conclusion of this chapter, the requirements for the CVFM's interface for image selection are revisited, the characteristics of the perceptually organised rectangular SOM browser are compared against them and it is chosen as the format to be used.

3.6 Conclusion to Chapter 3

The perceptually organised rectangular SOM browser has been chosen as the interface for image selection to be used to evaluate the CVFM. Table 3.5 revisits the requirements for the image selection interface (originally set out in Table 2.1) clarifying that, as far as is possible at this stage they have been met.

ISIR No	**Image Selection Interface Requirements (ISIR)** ***Details on whether the requirement has been satisfied?***
1	Intuitive image selection *The chosen format achieves this by* *a) Browsing* *b) Using perceptual data as opposed to computer vision features thus avoiding the" semantic gap"* c) *Using the self-organising map (SOM) format shown by Halley (2012) to outperform two other perceptually organised browsing environments.* *In addition Halley's method for obtaining perceptual similarity data enriched by crowds allows for large image sets (500 images) which can be further scaled.*
2	We must own the images or be allowed to use them. *The choice of the perceptually organised SOM browser will not affect this.*
3	Use a closed set of images *This is one aspect which makes the use of perceptual data possible due to the bounded nature of any image set deployed in the browser(see 3.3.4).*

Table 3.5 - Requirements for an image selection interface revisited.

This chapter began by setting out the requirements for the image selection interface. Content based image retrieval (CBIR) was introduced and the disadvantages of query-

based search were discussed. The advantages of browsing in relation to the intuitive requirement for the CVFM were exposed and so this approach was adopted. As data to structure the browser would be an important aspect of its success, possible sources of this data were discussed. Computer vision features, a common source of data for browsing and image search were introduced but it was noted that there exists a mismatch between the similarities of images as perceived by humans and as defined by computer vision features. This mismatch is known as the "semantic gap". The possibility of using perceptual data instead of features was examined and deemed possible due to the requirement of a closed set of images (ISIR3). An intuitive browsing environment developed by Halley (2012) based on perceptual data enriched by crowds and a rectangular self-organising map of image stacks was chosen. See Table 3.5.

Chapter 4

Constructing the Abstract500 SOM Browser

Returning to the thesis goal of *developing the means to implement the CVFM sufficiently to allow evaluation*, this chapter begins that development in a practical sense. A major conclusion from Chapter 2 was that images should be the medium for the new CVFM. Chapter 3 concluded that these images should be presented in a self-organising map browser based on perceptual data as developed by Halley (2012) and further investigated by Padilla et al. (2013). Thus, the purpose of this chapter is to describe the construction of an image set suitable for design feedback and its deployment in a self-organising map browser (creating the *Abstract500* SOM browser) to enable intuitive image selection.

In Section 4.1 requirements for the image set are formulated (see Table 4.1) with the aim of producing a component for the CVFM that would be flexible enough to enable the study of design feedback with images. Sections 4.2 and 4.3 discuss the type of image with which to populate the browser and the issue of image copyright. Section 4.4 describes how the images were gathered. Following that, Section 4.5 describes how the perceptual data required to organise the browser was obtained. Section 4.6 informally evaluates the perceptual data by a) dimensionality reduction and b) the creation of a 3D visualisation of the structure that the perceptual data brings to the image set. In Section 4.7 the Abstract500 image set is assembled in a SOM browser ready to be deployed in an evaluation. Finally, Section 4.8 concludes by revisiting the image set requirements and summarising the output of this chapter.

Appendix F is the appendix associated with this chapter.

Published work

The Abstract500 SOM browser features in all six publications listed in **Error! Reference source not found.**. In particular, Padilla et al. (2012) focuses on the work

described in this chapter. Also, Padilla et al. (2013), after describing the method used to produce two perceptually organised SOM image browsers (one being the subject of this chapter), proceeds to show their efficacy and flexibility as intuitive browsing interfaces.

4.1 Image Set Requirements

To establish a domain and parameters for the image set some requirements were formulated. Table 4.1 lists these requirements along with their motivation.

ISR No	Image Set Requirement (ISR)	Abbrev.
1	The images should be suitable for design feedback; i.e. designers should be familiar with the type of image from established design practice.	Suitable for design
2	The images should, if possible, avoid subjects which may bias feedback due to the images containing meanings specific to an individual user's life experiences which other users might not share, thus confounding communication.	Non-specific
3	The images should not contain recognisable symbols such as alphabetic, numeric or pictographic characters because the communication being investigated is to be outside the sphere of written language.	No symbols
4	Perceptual similarity data must be obtained on each image, to allow deployment in a SOM browser, thus permitting intuitive image selection.	Perceptual data
5	There must be a large number of images in the set to offer users a wide choice so as to avoid users feeling that their expression is limited by the visual "vocabulary". A pragmatic decision was taken to set the size to 500 a) allowing a wide choice b) defining the scope of the data collection task c) capping the processing overhead for the associated similarity matrix required by ISR 4.	Population 500
6	iPads may be used to display the browser so iPad screen size and resolution should be taken into account.	Resolution
7	The images must be free to use as a large number of images will be needed and negotiating licenced use of many proprietary images would be costly and time-consuming.	Free to use
8	The set should contain no duplicate images as this might confound later experiments.	No duplicates

Table 4.1 - Requirements for an image set with which to evaluate the CVFM. In addition to the numbering, abbreviations are included for reference.

4.2 Selecting the Type of Image

To address requirements ISR 1 and 2 (*Suitable for design* and *Non-specific*) a decision on what type of image to use was required. In Chapter 2 the use of images by designers

in mood boards was discussed and it was noted that abstract images are often used for this. It was also noted that a major reason for this was that abstract images have fewer specific figurative associations (Garner & McDonagh-Philp, 2001). This aspect of abstract imagery also fits with ISR 2, the requirement to avoid images that might hold some significance for one person and not another thus confounding communication. Thus abstract images can meet both ISR 1 and ISR 2 because a) designers are likely to already be comfortable with their use in mood boards and b) abstract images should have fewer specific figurative connections than, for example, portrait photographs, cityscapes, or landscapes.

Thus it was decided to seek abstract images for use in the browser.

4.3 Copyright

The use of Creative Commons licenced images was examined with ISR 7 (Free to use) in mind. As a minimum most Creative Commons licences allow an image to be used for non-commercial purposes as long as the owner is credited. Thus if, when gathering the images we a) restrict our search to Creative Commons licenced images and b) also gather attribution data and store it along with the images, it should be possible to achieve free use of a large number of images for the research.

Therefore it was decided to seek only Creative Commons licenced images and take steps to store attribution data along with the images.

4.4 Gathering the Images

Having decided, in 4.2 and 4.3, to seek Creative Commons abstract images, this section describes the practical steps taken to obtain a quantity of such images from the World Wide Web. It was decided to use a screen scraper application. Sections 4.4.1 to 4.4.3 set out some practical parameters for the screen scrape (Table 4.2), the requirements for a database with which to manage the images (Table 4.3) and rules for rejecting images as unsuitable, from those candidate images to be gathered (Table 4.4). Sections 4.4.4 and 4.4.5 describe an initial test screen scrape and then the full screen scrape gathering 1800 candidate images. Sections 4.4.6 and 4.4.7 set out how images were assessed for suitability for inclusion and how duplicates were eliminated. In Section 4.4.8 the final

500 images are allocated ID numbers for the *Abstract500* image set. Lastly, Section 4.4.9 summarises the outcome of this section.

4.4.1 Practical Parameters for the Image Screen Scrape

Practical parameters for the screen scrape were formulated and are summarised in Table 4.2. Details of the motivation for these parameters can be found in Appendix F p.231.

PP No	Practical Parameter
1	Source the images from Flickr
2	Gather 1800 images initially
3	Resolution 128x128 pixels minimum
4	Use Flickr "Safe Search"

Table 4.2 - Screen scrape practical parameters summary. The "PP No" column refers to the table in Appendix F p.231 which details the motivation for each parameter.

Thus images from Flickr, tagged with the word, "abstract", of at least 128x128 pixel resolution, recorded as Creative Commons free for non-commercial use, categorised as "safe" in Flickr safe search, were to be screen scraped.

4.4.2 Database to Manage the Images

Database requirements to allow the satisfaction of ISRs 1, 2, 3 (suitable for design, non-specific, no symbols), ISR 7 (free to use) and ISR 8 (no duplicates) were formulated (Table 4.3) and a database with which to manage the images was created.

DBR No	Database Requirement
1	Storage of attribution data
2	Image display
3	Image search by field
4	Assessment allowing images to be flagged as "Assessed" and "Suitable"
5	Allocation of Experiment ID
6	Fields to store image attributes to aid in the elimination of duplicates

Table 4.3 - Image management database requirements.The database had to allow these actions and have these facilities.

4.4.3 Rules for Accepting or Rejecting Candidate Images

Once gathered the candidate images would need to be assessed as suitable or rejected as unsuitable in relation to ISRs 1, 2, and 3 (suitable for design, i.e. abstract; non-specific; no symbols). While it might be possible to find computer vision features to recognise a

proportion of images containing alphabetic symbols or numbers (ISR 3), ISRs 1 and 2 were too subjective for a) definition precise enough to allow b) the current computer vision techniques to be applied so thoroughly that no further manual examination would be required. Thus, as there were only planned to be 500 images eventually in the set and a final manual check would be required anyway, it was decided not to expend resources on researching and applying computer vision techniques to algorithmically filter the images. The images would be manually viewed using the database management application and accepted or rejected (based on IRQs 1, 2, and 3). Also to be taken into account was IRQ 5 *"...image set to offer users a wide choice so as to avoid users feeling that their expression is limited..."*. Thus images which, although not exact duplicates, but were near duplicates would be rejected. In addition, presentation in the SOM browser meant that images which possessed a border intrinsic to the image would be unsuitable as this would affect the uniformity of presentation and, by attracting the viewer differentially to a bordered image, would affect the purpose of the browser.

Thus the rules in Table 4.4 for assessing and rejecting candidate images were formulated.

CIAR No	Candidate Image Assessment Rules (CIAR)	Ref ISR
1	No people	ISR 2
2	No full depictions of objects natural or man-made	ISR 2
3	No symbols or writing	ISR 3
4	No near duplicates	ISR 5, ISR 8
5	No Borders	

Table 4.4 - Candidate Image Assessment Rules along with their motivating ISRs.

4.4.4 Test Screen Scrape

A test screen scrape of 30 images (one page) fitting the parameters was done. The first 20 were taken as a sample and the Candidate Image Assessment Rules in Table 4.4 were applied. 15 out of 20 were accepted, thus confirming that the planned gathering of 1800 images would provide enough candidate images. The images in the sample scrape along with reasons for accepting/rejecting can be found in Appendix F p.232.

4.4.5 Screen Scrape

Thus 1800 images, tagged with the word, "abstract", of at least 128x128 pixel resolution, recorded as Creative Commons free for non-commercial use, and categorised

as “safe” in Flickr safe search, were screen scraped from Flickr. The scripts developed for the screen scraper recorded the data such as Flickr account name and the referring page URL (for attribution). These data included the URL to download the medium resolution version of the image. The medium resolution would be well in excess of the 128x128 resolution. They could be reduced later for use in the browser while still being available at this medium size (typically 600x450) if need be. The downloaded data was loaded into the abstract image database and the downloaded images collected for resizing and cropping.

The images were then resized and cropped to 128x128 pixel resolution by batch processing. See Appendix F p.233 for details.

4.4.6 Assessing Images for Suitability

By following the criteria in Table 4.4, 33% of the images in the pool were rejected. See Appendix F p.233 for details. (Section 4.4.4 refers to examples of rejection/acceptance during the test screen scrape.)

4.4.7 Elimination of Duplicate Images

ISR 8 (Table 4.1) requires there be no duplicates. Steps as were taken to identify duplicate images by sorting the database of images based on the average RGB values for the images. One instance of this was discovered and eliminated by rejecting the pair of images as unsuitable. See Appendix F p.234 for details.

4.4.8 The Final 500 Abstract Images

500 images, sampled from the images assessed as suitable, were allocated an experiment ID number and those 500 became the *Abstract500* image set.

4.4.9 Conclusion to Section 4.4

Section 4.4 described how the images for the *Abstract500 image set* were gathered. Practical parameters, database requirements, and rules for the rejection of images were established; a body of candidate abstract images was gathered and stored in a database along with attribution data. The unsuitable images and duplicates were identified and

rejected. 500 were sampled from those that remained establishing the *Abstract500 image set* of abstract images at 128 x 128 pixel resolution ready for assembling into a SOM browser.

However, perceptual similarity data would first be collected on the Abstract500 to satisfy ISR4 (perceptual data) (Table 4.1) and to enable SOM browser construction. The next section describes that.

4.5 Obtaining Perceptual Data on the *Abstract500*

Image Set Requirement No.4 (ISR4, Table 4.1) requires that perceptual similarity data be obtained on the images, to allow deployment in a SOM browser, thus permitting intuitive image selection. Thus the goal of this section is set out explicitly in Table 4.5 below.

Perceptual Data Requirement
The aim of the work described in this section is to produce a 500x500 similarity matrix describing the perceptual similarity (i.e. similarity as judged by humans) of each image in the Abstract500 to the other 499, thus satisfying Image Set Requirement No.4 (ISR4, Table 4.1).

Table 4.5 - Perceptual Data Requirement for the Abstract500.

The rest of this section starts in 4.5.1 with an overview of the method, developed by Halley (2012), to be used for obtaining the similarity data on the image set. In 4.5.2 and 4.5.3 the reasons for using crowdsourcing and for choosing Halley's method are set out. The approach taken to ensure the reliability of the crowdsourced judgements is described in 4.5.4 describing the necessary differences from the approach used by Halley. In 4.5.5 and 4.5.6 the conduct of the initial free sorting of a 'bootstrap' subset of the images and the crowdsourcing of the similarity judgements on the remainder of the Abstract500 image set is described. Finally in 4.5.7 the section concludes by revisiting the *Perceptual Data Requirement* and summarising how this it was satisfied.

4.5.1 Overview of the Method

The method used to obtain perceptual data on the Abstract500 image set is described in Padilla et al. (2013). In that work the method is termed "perceptual similarity enriched by crowds". The method is described in greater detail by Halley (2012, Chapter 10) who described the method as "data set augmentation".

This method established by Halley (2012) for obtaining human perceptual similarity data on an image set in a scalable way, can be summarised as involving two stages. The two stages are described here in the context of generating a similarity matrix to describe a set of 500 images, which is the number Halley used and is the same number as in the Abstract500 image set. Firstly, use free sorting of 100 reference images by lab participants to generate a 100x100 similarity matrix (termed the *bootstrap matrix*). Secondly, use remote crowdsourced participants to identify reference images (from the bootstrap matrix) that they view as 'similar' to the remaining 400 *query* images. The query images can then be added incrementally to the matrix. With each addition to the matrix, the new similarity vector is calculated as the average of chosen reference images' similarity vectors. The result is a 500x500 similarity matrix (termed the *augmented matrix*) which describes the perceptual similarity relationships between all the images in the set.

4.5.2 Why Crowdsourcing Is Used

Halley's work (2012) showed that for normal lab-based experiments reliable perceptual similarity data could only be obtained on up to 130 objects due to the fact that it is feasible for that number to be free grouped by a participant, in a single experimental session up to one hour long, without the data being affected by participant fatigue. To ask a participant to spend from two and a half hours to four hours free group 500 items is not feasible due to a) fatigue (both mental and physical) affecting data reliability b) difficulty recruiting participants committed to such a long task and c) ethical considerations in asking such effort of participants.

4.5.3 Why the Method Was Chosen

The crowd enrichment method for producing a large similarity matrix will be used in this thesis because a) its scalability allows for 500 images b) the scalability allows the set to be augmented later if required and c) it has been proven to produce an intuitive organisation both for monochrome texture images and for abstract images (Padilla. et al 2013).

In addition to these reasons for using this method, a further factor in favour of using it was that Halley had passed to the author code exemplars for a) the implementation of the crowdsourced augmentation, and b) the final browser assembly, which could be

adapted for the purpose of the work in this chapter (Halley 2011). Thus the cost of developing this code from scratch for this project would be avoided.

4.5.4 The Approach to Quality Control

Halley's method of obtaining scalable human perceptual similarity data on an image set is followed in this thesis for the Abstact500 with one exception. That exception is in one aspect of the approach to quality control of crowdsourced participants' observations. This exception is described in this sub-section by comparison with Halley's method.

A specific issue in employing crowdsourced participants in providing judgements is that of "cheaters"; i.e. avoiding accepting into the data, judgements from insincere participants who seek to exploit the crowdsourcing platform for unfair monetary gain.

The Amazon Mechanical Turk (AMT) (2015) crowdsourcing service used by Halley (2012) permits those who commission workers to do tasks to offer a bonus for superior work. It is also possible to deny payment for poor quality work not sincerely attempted. Use of AMT was also available for the work in this chapter of this thesis and so AMT would be used.

These facilities, available through AMT, allowed what might be termed a "carrot and stick" approach to quality control; i.e. a bonus could be offered for good work and payment could be denied for insincere work. AMT workers have an added incentive (other than not getting paid) not to claim for poor work as workers whose claims are rejected suffer a reduction in their recorded level of work accepted and this can eventually affect the availability to them of tasks for which the qualification is a minimum level of past work acceptance (Kosara, 2010).

Halley offered a bonus to AMT workers who provided additional data over and above the minimum. As this was also possible for our data gathering task we would offer a bonus and on a similar basis.

Where our approach would by necessity differ from Halley's would be in how work was to be assessed as not a sincere attempt and so rejected.

Halley (2012) used a "gold set" approach to quality control (Kazai, 2011) when it came to accepting or rejecting the data from crowdsourced participants; i.e. a portion of each participant's judgements are sought on items for which data is already known (so called

gold data). The reliability of a participant's observations is then estimated by the veracity of their judgments on the gold data items and the totality of their observations is either accepted or rejected on that basis. Halley (2012) was able to use this approach as that work also involved a comparative study and had already produced other reliable perceptual data on the images being investigated.

However, no such prior reliable data existed on the Abstract 500 image set. Therefore a different approach was adopted. The approach used *time on task* to flag up individual results sets which might represent ill-considered and hurried observations submitted only to claim payment. Borderline cases could be scrutinised manually to avoid the unnecessary discarding of acceptable data. See Appendix F p.234 for details. On this basis participant's observations would either be accepted or rejected.

However, before any crowdsourced observations could be sought, a bootstrap browser would need to be constructed based on lab-based participant free sorting of 100 images from the Abstract500. This is described in the next sub-sections.

4.5.5 The Bootstrap Sort

To provide a scaffold image set structure for crowdsourced workers to use as the reference or bootstrap from which to draw likenesses for the majority of the Abstract500, Halley's method required that a subset (100) of the Abstract500 be free sorted by lab-based participants forming the reference or bootstrap similarity matrix. Thus 100 of the Abstract500 were free sorted by 20 participants (11 male) in the lab. (Figure 4.1). See Appendix F p236. for details.

Figure 4.1 - A participant free sorting the bootstrap subset images (100).

The 20 sets of perceptual groupings produced by the participants in the free sorting resulted in a 100x100 similarity matrix in which the similarity between any two images is the frequency with which that pair of images were grouped together by the participants normalised by dividing by the number of participants.

4.5.6 The Crowdsourced Augmentation of the Matrix

The 100x100 bootstrap similarity matrix was input to the SOM toolbox for MATLAB (Vesanto et al., 1999) and the resulting SOM structure used to inform the construction of a bootstrap SOM browser as described by Halley (2012) and in Padilla et al (2013). This bootstrap SOM was implemented in the image set augmentation interface for presenting to the AMT participants (Figure 4.2). The image IDs for the remaining 400 images in the Abstract500 were packaged into randomly formed stimuli packets to be served in batches of 20 query images per crowdsourced participant. The augmentation of the 100x100 bootstrap matrix with the 400 remaining images to form the 500x500 Abstract500 similarity matrix was carried out as described by Halley (2012) and Padilla et al. (2013). See Appendix F p236 for details.

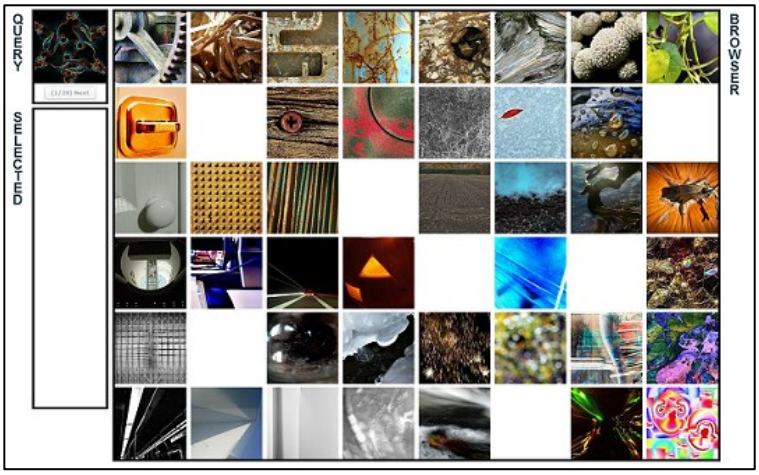


Figure 4.2 - Image set augmentation interface for AMT

The resulting 500x500 similarity matrix describing the similarity of each image provided a convincing organisational structure for the Abstract500 image set as shown in the next section (4.6).

4.5.7 Conclusion to Section 4.5

This section commenced with an explicit statement (in Table 4.1) of the *Perceptual Data Requirement*. This is revisited below in Table 4.7 to clarify that it has been met.

Perceptual Data Requirement ***Has this been satisfied?***
The aim of the work described in this section is to produce a 500x500 similarity matrix describing the perceptual similarity (i.e. similarity as judged by humans) of each image in the Abstract500 to the other 499. Thus satisfying Image Set Requirement No.4 (ISR4, Table 4.1). *A 500x500 similarity matrix of human perceptual similarity judgements describing the Abstract500 image set has been created.*

Table 4.6 - Revisiting the Perceptual Data Requirement for the Abstract500.

This section began with an overview of Halley's (2012) method of obtaining scalable large perceptual similarity matrices, and set out why crowdsourcing and the method should be used in this case. The need for an alternative approach to quality control of the data from the crowdsourced participants was discussed and an alternative approach was set out. The initial free sorting of a subset of the images and the actual crowdsourcing of the similarity data on the rest of the image set was described.

The resulting output of this section is the 500x500 similarity matrix describing the Abstract500 image set. In the next section that data is informally evaluated.

4.6 Evaluating the Perceptual Data Using MDS

The fact that the Abstract500 image set is described by a 500x500 similarity matrix means that, in theory, there could be up to 500 dimensions in the data. In practice it is likely that there would be less than 500 dimensions. For example possible dimensions might include the red/green/blue colour dimensions. One way of appreciating the dimensionality of multivariate data is by using multidimensional scaling (MDS) (Cox & Cox, 2001).

Following methods from Martinez et al. (2011) classical MDS (which can be considered similar to a principal coordinate analysis (Cox & Cox, 2001)) was applied to a dissimilarity matrix calculated from the Abstract500 similarity matrix (See Equation (4.1)).

$$Dissimilarity\ Matrix\ = 1 - Similarity\ Matrix \qquad (4.1)$$

According to Martinez et al. (2011) the Eigenvalues produced from the MDS provide information on the dimensionality of the data being explored. A scree plot of the Eigenvalues and their indices (which represent the data dimensions) reveals the actual dimensionality within the data by illustrating the Eigenvalue index (or dimension) at which an "elbow" occurs. Figure 4.3 shows the scree plot for the MDS of the Abstract500 data. The chart shows that the data are indeed multidimensional but the Eigenvalues begin to level out in relation to each other between dimension numbers 12 to 20 indicating that there may be up to 20 significant dimensions to the data. (Note: later, in Chapter 6, alternative methods of dimensionality reduction, including non-metric MDS, were applied to the data. These suggested it may have lower dimensionality than indicated by the classical MDS).

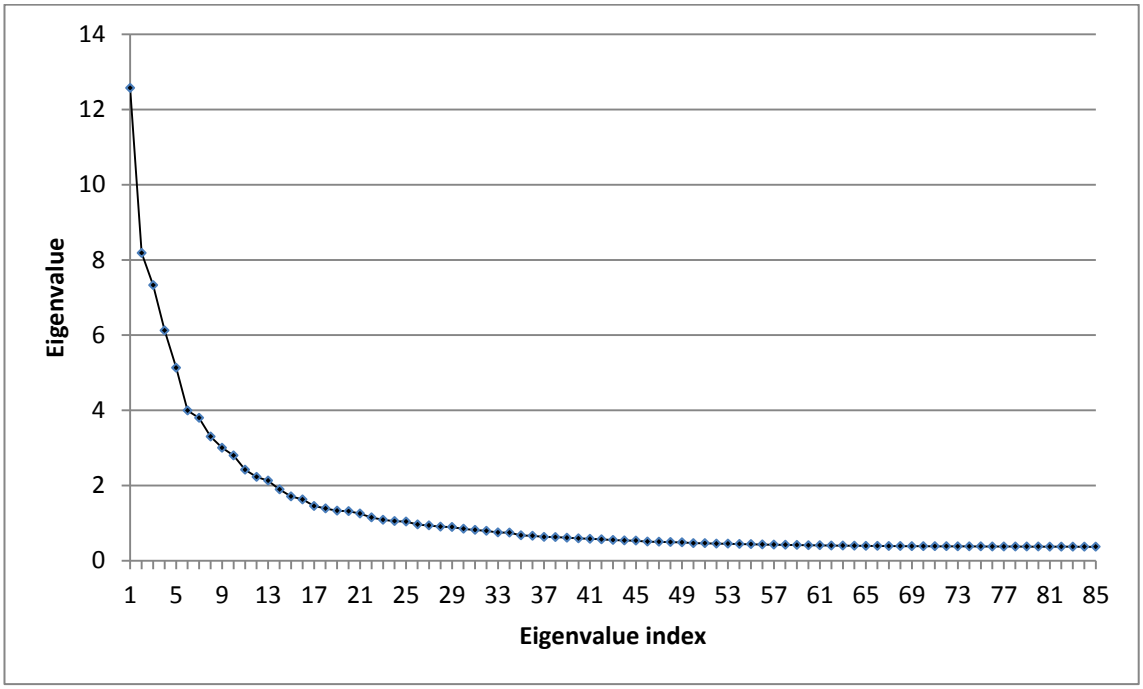


Figure 4.3 - Scree plot of Eigenvalues from classical MDS of the Abstract500 dissimilarity matrix. The Eigenvalue index corresponds to the dimension (or coordinate) number.

The SOM browser is one way of visualising the structure in the data describing the Abstract500 and that is shown in the section following this. However, an alternative 3D visualisation of the data was created based on the 3 most significant dimensions from the classical MDS analysis of the Abstract500 data matrix (using a method developed by Halley (2012) during a comparison of browsing environments, and based on work by Rogowitz et al. (1998)). The dimensionality reduced structure of the data was viewable in a 3D rotatable space.

Screenshots of the different rotated aspects of the 3D MDS view and selected regions from it can be seen in Figure 4.4 below and in Appendix F p239.

An informal evaluation of the organisation within the image set based on the similarity matrix began by examining the 3D MDS view. It could be seen that there were regions and clusters clearly representing themes within the image set; e.g. there was a structural themed cluster consisting of unusual architectural views, a natural themed cluster of various unusual views of plants, and a cluster of highly coloured classically abstract patterns.

Figure 4.4- Classical MDS 3D view. Screenshot of one aspect. Further views are in Appendix F p239.

The informal evaluation of the 3D MDS view of the *Abstract500* image set included showing it to a small number of staff and students of the University's School of Textiles and Design who might be taken as representative of designers (possible future users of a system of visual feedback). All were engaged, indeed fascinated, both by the image collection and its structure when exploring the *Abstract500* image set in the 3D view.

From this informal evaluation using 3D MDS and a rotatable visualisation it was concluded that

a) the crowdsourced similarity matrix augmentation had worked in that it had produce a sensible structure for the *Abstract500* image set (at least in the 3 most significant dimensions) and

b) the *Abstract500* image set and its perceptual structure in 3D was appealing to designers during the informal evaluation.

4.7 Assembling the Abstract500 SOM Browser

By adapting the exemplar code passed to the author by Halley (2011) the *Abstract500* was assembled into an 8x6 stack SOM browser. (8x6 was used as it was planned to deploy it on iPads for an experiment later and this configuration could be accommodated on an iPad display.) The browser is shown in Figure 4.5.

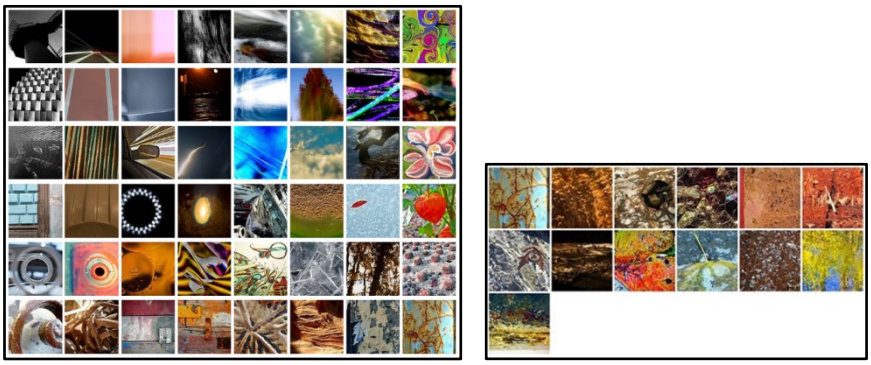


Figure 4.5 –A rectangular SOM browser presenting a large image set in 8x6 stack configuration. Top level (left) and the bottom right hand corner stack opened (right).

Each stack in the browser represents a cluster of images based on the similarity matrix. The image nearest the centroid of the cluster (by Euclidean distance) (*the centroid image)* is that chosen as the top image in the stack; i.e. the image which appears on the top level of the SOM to represent the stack is the centroid image. Tapping or clicking that image opens the stack. The images are listed within the stack by Euclidean distance from the centroid, lowest first. Adjacent stacks contain images that are similar. Stacks far apart contain dissimilar images. Here "similar" and "dissimilar" are objectively defined by the collective similarity judgments of the lab participants who did the bootstrapping free sort and crowdsourced participants who likened their query images to those in the bootstrap browser.

The stacks in the resulting SOM contained sensible themed subsets of the Abstract500 which, on an informal basis, as with the 3D view, when showing it to staff and students at the University's School of Textiles and Design, was found to be captivating and engaging to explore. Deployed on an iPad it had even more appeal as this introduced the touch interaction with the SOM.

4.8 Conclusion to Chapter 4

The aim of this chapter was to create a perceptually organised SOM browser populated with images suitable for use in the evaluation of the CVFM. Requirements for the images destined for use in the browser were set out in Table 4.1. These requirements are now revisited in Table 4.7 to establish that they have been satisfied.

ISR No	Image Set Requirement (ISR) *Details on whether the requirement has been satisfied?*	Satisfied Yes/No
1	The images should be suitable for design feedback; i.e. designers should be familiar with the type of image from established design practice. *Abstract images were sought (a type already used by designers in mood boards). Informally, creative people find the image set engaging.*	Yes
2	The images should, if possible, avoid subjects which may bias feedback due to the images containing meanings specific to an individual user's life experiences which other users might not share, thus confounding communication. *Steps were taken to gather only images associated with the term "abstract". Rules were established to reject images that were too figurative (Table 4.4) and such images were rejected.*	Yes
3	The images should not contain recognisable symbols such as alphabetic, numeric or pictographic characters because the communication being investigated is to be outside the sphere of written language. *Such images were rejected from those gathered.*	Yes
4	Perceptual similarity data must be obtained on each image, to allow deployment in a SOM browser, thus permitting intuitive image selection. *A 500x500 perceptual similarity matrix was created by following Halley's (2012) crowdsourced matrix augmentation method.*	Yes
5	There must be a large number of images in the set to offer users a wide choice so as to avoid them feeling that their expression is limited by the visual "vocabulary". A pragmatic decision was taken to set the size to 500 a) allowing a wide choice b) defining the scope of the data collection task c) capping the processing overhead for the associated similarity matrix required by ISR 4. *The Abstract500 contains 500 diverse abstract images*.	Yes
6	iPads may be used to display the browser so iPad screen size and resolution should be taken into account. *The images are 128x128 pixel resolution and an 8x6 stack SOM presentation of the Abstract500 will fit on a iPad1 display.*	Yes
7	The images must be free to use as a large number of images will be needed and negotiating licenced use of many proprietary images would be costly and time-consuming. *The Abstract500 images are all Creative Commons licenced as free for non-commercial use and an associated database keyed by image ID holds attribution data on each image.*	Yes
8	The set should contain no duplicate images as this might confound later experiments. *Duplicates were eliminated by analysing each image's mean RGB, storing it in the database and sorting the images based on this data such that duplicates appeared side-by-side. One pair of duplicates was found and rejected from the database.*	Yes

Table 4.7 - Revisiting the requirements for a image set to evaluate the CVFM.

Thus the requirements set out for the image set in the introduction to the chapter have been met. Abstract images were chosen as a suitable type with which to populate the browser and Creative Commons licenced abstract images were gathered from Flickr (attribution data being stored in a database). The images were examined. Duplicates and those not suitably abstract (based on rules set out in Table 4.4) were rejected. 500 were sampled from those that remained forming the *Abstract500 image set*. Following a method developed by Halley (2012) and reported in Padilla et al. (2013) perceptual similarity data describing the image set was gathered through lab-based free sorting and crowdsourced grouping. The resulting 500x500 similarity matrix was informally evaluated by producing a 3D visualisation based on the three most significant dimensions established in a dimensionality reduction analysis (classical MDS). The 3D visualisation illustrated sensible structure within the data and was found, informally, to be engaging for creative people. The similarity matrix was used to inform the construction of an 8x6 stack rectangular SOM browser presentation of the Abstract500. Deployed on an iPad this SOM browser was also, informally, found to be fascinating for staff and students at the University's School of Textiles and Design.

Thus one component, the Abstract500 SOM browser, for evaluating the CVFM was developed.

Chapter 5

Image Summarisation

Restating the thesis goal of *developing the means to implement the CVFM sufficiently to allow evaluation*, the previous chapter produced the Abstract500 browser, an intuitive perceptually organised browser containing abstract images as one component to enable evaluation of the CVFM. Referring to Figure 1.1, the Abstract500 browser will be used by *crowd users* (individual members of the crowd) to express their reaction to a design shown to them by a *designer user*. That reaction will consist of a number of images selected from the browser. These images will be collected along with those from other crowd users. The CVFM requires that these gathered images be summarised into a concise visual summary to be shown to the designer who in turn will form an impression of the crowd's reaction to the design. Thus image summarisation is the other major component required for evaluation of the CVFM.

The purpose of this chapter is to investigate visual summarisation to decide on an approach to be used in this thesis. First, in 5.1 and 5.2 the need for summarisation, (initially noted in 2.4.2), is recapitulated and the requirements for image summarisation are set out. In Section 5.3 approaches to image search at scale on the World Wide Web and summarisation and of images from social media are discussed. Then in 5.4 existing work on summarisation of defined image sets is examined and its purposes, aspects and approaches are identified. Section 5.5 criticises these methods for summarising defined image sets against one which would use the reliable similarity data for the Abstract500 image set. Section 5.6 in Table 5.2 compares the existing methods with a hypothetical ideal method for the situation at hand, and then suggests that a method specific to this situation be developed. Finally, in 5.7 the chapter is summarised concluding with a recommendation that a method of summarisation specifically for the CVFM be developed but also identifies one of the existing methods as a possible alternative to be adapted should any difficulty arise in that development.

5.1 The Need for Image Summarisation

In 2.4.2 it was pointed out that for use to be made of the value in a crowd's collective judgement, the judgements of the individuals in that crowd collectively must be summarised to be read conveniently. 2.4.2 also concluded that, for the CVFM, with each crowd member contributing *images*, the challenge will be to summarise the totality of the crowd's *image choices* into a meaningful but concise form.

The crowd could be large. Indeed a successful deployment of the CVFM would involve large amounts of data if the aim of engaging a crowd as a potential customer base was to be achieved.

Thus, a method of summarising selections made from the abstract image browser is required to facilitate convenient consumption of the feedback by designer users.

5.2 Requirements for Image Summarisation Method

Requirements for the summarisation method are set out in Table 5.1 with motivations.

SMR No	Summarisation Method Requirements (SMR)	Motivation
1	Exploit existing perceptual data	The structure in the Abstract500 perceptual data was demonstrated to make sense during our informal evaluation using a 3D classical MDS visualisation and was engaging for creative people (See 4.6).
2	Non-overlapping image placement	Later, when formally evaluating the semantic performance of the summaries, should some part of an image be obscured this could have an effect on the semantic content or emotive impact of the image and therefore on the results of any evaluation.
3	Be designed to cope with collections such as the Abstract500 e.g. lack of faces	The Abstract500, while diverse, is not typical of image sets which most summarisation methods are designed for; i.e. photos of places and people.

Table 5.1 - Requirements for an image selection interface with motivation for each.

5.3 Image Search and Summarisation at Scale

In this section methods applied to clustering and ranking of images at very large scale for web search (5.3.1) and social media image summarisation (5.3.2) are described. Although the context and scale of these applications is quite different from the image sets and application that are the focus of this thesis they are relevant in providing

context to the more focused examination of summarisation work on closed image sets described in the section which then follows this one.

5.3.1 Search

Text labels

Luo et al. (2011) state that "commercial search engines and Web albums rely on text annotations associated with images for indexing and retrieval tasks". One source of information used in labelling images from the World Wide Web for retrieval is the text associated with them on a web page. However, because this can lead to mislabelling of images and thus erroneous search results (Von Ahn & Dabbish, 2004), from 2006 until 2011 a game called "ESP" was used by Google as "Image Labeler" (Von Ahn & Dabbish, 2004). The game was an effort to improve the labelling of Google's indexed images based on crowdsourcing of labels. Players were motivated to describe an image with words that coincided with another player's description of the same image. This generated labels which were more semantically reliable than the text found proximate to the image on a web page. The game had gathered 50 million labels by 2008 (Von Ahn & Dabbish, 2008) and continued in popularity. However, in 2011 Google withdrew it. The official blog post informing of the withdrawal of "Image Labeler" did not give a reason for the cessation (Google 2011). The author is left to speculate that Google had calculated that relying on the game crowd could not meet the labelling capacity required to index all web images and that a fully automated solution would have to be deployed.

Reranking Search Results from Text Label Based Search

To improve on the often noisy results from image search based on text labels work has been done on reranking the results lists from such searches by applying additional techniques to the initial results lists (which offers less of a computational challenge as the methods are being applied to smaller finite subset of the images being searched). In a recent review of these techniques, Mei et al., (2014) identified four categories of methods.

Self-Reranking,

These methods use only the information within the ranked list returned by the initial search based on text labels. There are further subcategories:

- Clustering-based reranking: In principle clustering should serve to separate the more relevant results from those less so.
- Pseudo-relevance feedback. This assumes that the results already ranked highly from the label based search are the most relevant and these are used to train classifiers which then classify the remaining results.
- Object recognition-based reranking. Computer vision object recognition techniques are used to calculate the similarities between the search results.
- Graph-based reranking. This is based on the PageRank method (Brin & Page, 1998) effectively making use of the hyperlinked web structure associated with each of the results to inform the reranking. Alternative graph representations can be used such as Random Walk (Hsu & Chang, 2007).

Example-based reranking:

The user provides some query example images to accompany their text query. These examples are used to train classifiers.

Crowd reranking Methods:

These make use of search results from several search engines rather than just that one producing the initial results to be reranked. Having gathered these alternative results, common patterns can be derived and applied to rerank the initial results.

Interactive reranking:

Input from the user by way of annotation or rejection of a portion of the initial search results informs the reranking of the whole returned results list.

Thus it can be seen that a number of methods have been used to improve on world wide web image search retrievals based on text labels associated with the images. Some of these reranking methods, while improving web search, make use of data (such as the graph based reranking) which would not be available in the closed image sets used in this thesis. Also the computer vision techniques which could be used suffer from the semantic gap problem described in 3.3.2 when used on their own.

5.3.2 Summarisation of Social Media Images

Also dealing with images at large scale, work has been done recently on summarising images from social media on a given topic. That work (McParlane, et al., 2014)

focusses on summarising images of events. The challenges faced to summarise an event in images included dealing with irrelevant images (images with captions known as memes, screenshots, reaction or emoticon-style images not actually depicting the event), and also duplicate and near duplicate image detection. The images were sourced from tweets on a microblogging site (i.e. Twitter (2015)) and from Internet URLs posted in the tweets. The tweets from a defined one-month time frame were clustered into 50 separate events using the Stanford Parser (Klein & Manning, 2003).

As images from linked web sites were also used, additional irrelevant images were associated with these and some initial filtering steps, such as on filename to eliminate logo images and on dimensions to eliminate standard advert banners, were done.

The near duplicate image detection was done using a hashing function method (Tang et al., 2012). As hashing an image in this way produces a short string descriptive of the image this allows detection of near duplicates with a low processing overhead. The Perceptual Hash method (Tang et al., 2012) has good performance in detection of near duplicate images which have been transformed by resizing, cropping and exposure manipulation. The hash string for each image was calculated and the hamming distance (the number of bits which differ) is taken as the similarity measure between two images. The resulting clusters allow only one image from a cluster to be selected as representing the other near duplicates.

With near duplicate images eliminated the irrelevant images were next tackled. The screenshots and reaction or emoticon style images are computer generated and this category of image can be detected by using a classification model (Wang & Kan, 2006) to train a Support Vector Machine classifier. Colour histogram and edge histogram data (Manjunath, et al. 2002) were extracted from the model images to train the classifier and from the twitter images for classification following that. Detecting the meme images required the authors use a different approach due to multiple captions on any given meme background image. Therefore a local feature matching, using the SIFT feature set (Lowe 2004), was used. The meme background images from an archive were analysed for their SIFT features and then the same was done for the tweet images allowing matching of the tweet images with the archive of meme backgrounds. In this way the memes were removed.

Thus filtering out the unwanted images from the desired social media event images was far from trivial. What remained to be done was the ranking of the images to detect the

most relevant images whilst maintaining diversity within the top ranked images so as to achieve a selection of images which provided an overview of the event. An image's popularity within the tweets was a factor employed. Spam images injected into the tweets by spam bots although also at high popularity levels were popular across the tweets for all events and this could be detected and the spam images eliminated. Thus high popularity of an image led to its ranking highly relevant. However, to avoid a single scene dominating an event summary semantic clustering of the tweets containing the images (based on image content and time clustering (McMinn, et al. (2013)) was used and high ranked images from within separate clusters were selected. The authors evaluated the summary image presentations versus text and word cloud presentations and found that crowdsourced participants found that the image presentation helped them to understand the events depicted better that text and word clouds and also found them more engaging.

Thus the work described above shows that summarisation of social media images at scale is becoming possible. However, in the context of the image sets being deployed in this thesis, different challenges are faced. Rather than relevance to a given topic and elimination of spam, memes and near duplicates within a very large body of images, the challenges are more focussed on a closed set of images and summarising selections made from within those. Thus in the next section several works on summarising images in defined image sets is examined.

5.4 Summarising Defined Image Collections

Previous work in the area of summarising defined image collections (i.e. those not on the scale of the World Wide Web) addresses the problems encountered in the application of browsing very large (1000s) image collections (Fan et al., 2008). The other application is in producing summary collages as overviews or front pieces or introductory photo collage pages to precede or introduce more defined image collections; i.e. to produce summary collages for small (10s) to large (100s) image collections (Egorova et al., 2008) (Rother et al., 2006) (Lee et al., 2010) (Tan et al., 2011) (Xu et al., 2011) and to produce summary collages from a discrete handful of images without the need for reduction in image numbers (Favorskaya et al., 2012) (Wang et al., 2006). Additionally one application (Ahern et al., 2007) requires representative images be chosen, from those available, for particular geographical locations for the purpose of illustrating an interactive map.

An examination of the work in this area has thus revealed the purposes, aspects and approaches summarised in 5.4.1 and 5.4.2.

5.4.1 The Purposes of Summarisation

The existing summarisation methods have been developed for the following purposes:

1) Browsing very large (1000s) image collections.
2) Producing summary collages for small (10s) to large (100s) image collections.
3) Producing summary collages from a discrete handful of images without the need for reduction to a small number of representative images.
4) Choice of representative images to illustrate geographic locations.

5.4.2 The Two Aspects of Summarisation

The existing summarisation methods reveal two aspects to summarisation:

1) Reduction from many images to a small number of representative images.
2) Placement of the representative images on the summary.

5.4.3 Approaches to Reduction

One approach defines representative images as images that are interesting but different. Images are ranked by importance based on computer vision techniques such as face detection (Lee et al., 2010). The choice of high ranking candidate images is then filtered so as to rule out near duplicate images by using similarity data based on colour histogram techniques such as the hybrid graph representation (Park et al., 1999). Tan et al. (2011) clustered on similarity using computer vision techniques e.g. colour histogram.

Another approach to the reduction is by clustering the images and choosing representative images based on the cluster structure. Egorova et al. (2008) use source and date/time metadata as the data for clustering. Fan et al. (2008), in their work to improve browsing in large image collections, used tags already associated with the images to allocate images to topics. Within topics the images were then clustered based on similarity data calculated from colour, texture and interest point features. Xu et al (2011) adopting a related approach in that they make use of tags, termed their method,

"Hybrid image summarization". This relies on each image being accompanied by an associated text, treating the component words as tags and calculating similarity based on the tags in addition to similarity based on features.

In summary there are two main approaches, one uses importance ranking, and the other uses clustering. The sources of data for these approaches can be computer vision features, or tags (folksonomy, of context-based, or even time and location).

5.4.4 Approaches to Image Placement

There are two main approaches to image placement. The most common which we will term *packing*, takes account of the number of images, size of regions of interest within images, and orientation, to optimise use of canvas space. The other approach involves placement according to structure of relationships within the image set and seeking to preserve these spatially while projecting them onto the canvas space.

There is an additional issue: that of overlapping images in the summary and/or blending or blurring the boundaries between them to produce an artistic collage-style effect.

5.5 Criticisms of the Existing Methods

Despite the title of this section it should be noted that the methods described in the preceding sections are all valid approaches to the problem of summarising often large and fluid image collections. The criticisms in this section are made from the point of view of already possessing reliable perceptual similarity data on the images to be summarised (c.f. 4.8).

The existing methods all rely to some extent on computer vision techniques to measure the similarity between images. In 3.3.3, when considering the type of data for structuring the image browsing for the CVFM, the mismatch between similarity based on computer vision features and that based on actual perception led to the conclusion that perceptual data was more reliable than features.

Indeed, the summarisation methods which use metadata and tags are seeking to address the semantic gap between what can be deduced about the meaning of the image from its features and what the image actually means to a viewer. While it would be possible to harvest the tags associated with the Abstract500 images from Flickr, folksonomy tags

can be unreliable (Lee & Yong, 2008). (Indeed our own search based on the term, "abstract", yielded images many of which could in no way be described as abstract). The time/date metadata (Egorova et al., 2008) may contain irrelevant coincidences as the images gathered for the Abstract500 were made by many different Flickr users.

Thus existing methods of choosing representative images from image collections suffer from some drawbacks bearing in mind we possess the reliable similarity data on the Abstract500.

5.6 Overview of Methods and an Ideal Method

Paper Short Name (Reference)	Representative Image Choice						Image Placement Method				
				Data basis							
	Reduction	Clustering	Importance Ranking	Features	Tags (Folksonomy or Contextual)	Perceptual	Packing	ROI\Saliency	Features Similarity	Perceptual Similarity	Overlapping/blending
IDEAL FOR CVFM	✓	✓				✓				✓	**X**
Image Hive, (Tan et al., 2011)	✓	✓		✓			✓		✓		✓
Hybrid (Xu et al., 2011)	✓	✓		✓	✓		N/A	N/A	N/A	N/A	N/A
Geographic (Ahern et al., 2007)		✓			✓		N/A	N/A	N/A	N/A	N/A
Semantic (Fan et al., 2008)	✓	✓		✓	✓				✓		✓
Picture Collage (Wang et al., 2006)	N/A	N/A	N/A	N/A	N/A	N/A		✓			✓
Intelligent Collage (Favorskaya et al., 2012)	N/A	N/A	N/A	N/A	N/A	N/A	✓	✓			✓
AutoCollage (Rother et al., 2006)	✓	✓		✓			✓	✓			✓
Mobile Photo Collage (Lee et al., 2010)	✓		✓	✓			✓	✓			✓
Collage for PhotoBook (Egorova et al., 2008)	✓	✓			✓[3]		✓				✓

Table 5.2 - Comparison of existing methods with a hypothetical ideal method.

[3] The contextual data here is time data. Egorova et al (2008) acknowledge it does not work well for images taken on different cameras. i.e. it is designed for use on one person's photo collection.

Table 5.2 sets out the existing methods in comparison with a hypothetical ideal method for the CVFM. This ideal method is motivated by the following factors:

a) We do, already possess reliable, perceptual similarity data on our *Abstract500* set, gathered to construct the intuitive browser. This can be used to cluster image selections made from the Abstract500 thus enabling the calculations to choose representative images without having to resort to less reliable computer features. See SMR1 (Table 5.1).

b) The informal evaluation of the structure in the Abstract500 perceptual data was demonstrated to make intuitive sense during our evaluation using a 3D classical MDS visualisation (See 4.6). If these spatial relationships could be successfully preserved in 2D then we should aim to use the perceptual data to inform image placement on the summaries as well as for reduction to choose the representative images. Again see SMR1 (Table 5.1).

c) SMR2 (Table 5.1) precludes image overlap and anything that will obscure any part of an image so as to avoid confounding any evaluation experiments.

Notable points from Table 5.2 are:

1) All the existing methods involve overlap/blending of images and so are unsuitable

2) None make use of perceptual data.

3) The method of Tan et al. (2012) comes closest to our requirement.

4) To meet the ideal specification a summarisation method specific to the CVFM should be developed.

5.7 Conclusion to Chapter 5

In this chapter the need for image summarisation to condense crowd user images down to a concise summary for designer users was identified. Requirements for a summarisation method were formulated motivated by already having an image set with reliable perceptual similarity data. Existing work on image summarisation was examined; its purposes, aspects and approaches were enumerated, and described. In Table 5.2 the existing methods were compared with a hypothetical ideal method.

The comparison revealed that none of the existing methods is ideal. Thus a summarisation method specifically for the CVFM should be developed with both reduction to representative images and image placement based on the perceptual similarity data already in existence. Should some obstacle prevent such development then the existing method closest to the ideal is that of Tan et al. (2011) and might be adapted in that circumstance.

Chapter 6

Development of an Algorithm for Image Summarisation

Having concluded in Chapter 5 with the requirement that an image set summarisation method (based on clustering and the perceptual data we already possess) should be developed, the purpose of this chapter is to describe that development process. Table 6.1 and the subsequent text describe the explicit requirements for the algorithm.

SAR No	Summarisation Algorithm Requirements (SAR)	Motivation
1	Inputs: a) Image selection lists (ISLs). Such lists comprising sequences of image IDs as chosen from given image sets as responses by participants; b) feature or similarity vectors describing each image. These should be exploited to inform clustering and image position on summaries. Output: A non-overlapping montage of *k* representative images each sized proportionate to the number of image selection each represents.	The ISLs are the raw input for summarisation. See 6.6 for the rationale for non-overlapping images on summaries. See 6.8.2 for the considerations for setting the value for *k* in the experiments in this thesis.
2	An ISL can and probably will contain multiple occurrences of any particular image ID. Such repetition should be reflected in the weighting of that image ID in both a) the calculation of any cluster centroid representative image and b) be reflected in the presented size of the image representing the associated cluster.	This will allow the summaries to reflect popular image choices by having each cluster centroid drawn towards image choices which occur multiple times within the clusters. Thus it will be more likely that a popular image choice in the ISL will become one of the representative images on the summary.
3	The popularity of the images included in any given cluster in a summary should be reflected in the size of the image representing that associated cluster.	This is to provide a visual cue as to the relative population of each cluster when a summary is viewed.
4	The minimum length of an ISL (the minimum number of image selections) should be >= k. (Where k is the number of clusters and the number of representative images desired on the summaries). There need be no maximum length of an ISL for the purposes of the experiments in this thesis as the numbers will not be large i.e. <100. (However, see accompanying text below.)	It will not be possible to cluster any less than k images into k clusters.

Table 6.1 - Requirements for the summarisation algorithm with motivation for each.

There need be no maximum length of an ISL. However where this exceeds the number where the computational power available for the *k*-means clustering becomes an issue (e.g. 5000) then the population of each image ID within the ISL can be divided by the lowest common denominator for all the individual image ID populations within the ISL. Thus each image ID within the ISL will still be effectively weighted by its population proportional to the others. For the purpose of the experiments in this thesis, as the numbers were low, this step was not implemented. Note that, for larger implementations, the point at which the number of items being clustered becomes an issue will vary depending on the k-means clustering implementation that is deployed.

The remainder of this Chapter is organised thus: Section 6.1 sets out an overview of the summarisation method. Sections 6.2 and 6.3 describe considerations of the methods for clustering and dimensionality reduction. (3D visualisations of the Abstract500 image set are used to illustrate the choices of dimensionality reduction methods and these can be found in Appendix A. Section 6.4 describes why a two-stage dimensionality reduction is used. That section also relates the decision to exploit MDS of the perceptual data as the source of the positioning data for the component images on the visual summaries. Sections 6.5 to 6.7 set out the additional components of the overall algorithm developed to carry out the summarisation. Section 6.8 describes the implementation of the summarisation which is later used to evaluate the performance of the algorithm. The chapter concludes with a summary of the algorithm and a critique of the implementation.

Appendix A is the appendix associated with this chapter.

Published Work

The summarisation algorithm, or visual summaries made using it, feature in published work: Robb et al. (2015a), Robb et al. (2015b), Kalkreuter et al. (2013) and Kalkreuter and Robb (2012).

6.1 Overview of Planned Summarisation Method

Figure 6.1 shows an overview of what the summarisation should do. The selection of images to be summarised can, indeed probably will, include images selected multiple

times, and the summarisation will take that into account, giving more weight to images that were chosen repeatedly.

Figure 6.1 - Overview flow diagram of planned summarisation method.

To summarise selections from the Abstract500 image set, both the choice of representative images, and the placement will be informed by the perceptual data we hold on the image set. However, the summarisation method we intend to develop should be able to be applied to selections from any image set on which perceptual data exists.

As discussed in Chapter 5, the choice of representative images will be informed by clustering based on the perceptual data and their placement in a 2D summary will also be informed by the structure within the perceptual data.

Why 2D and not 3D?

There are two factors which favour a 2D summary over a 3D summary:

a) A 3D representation would require interactivity in navigating in 2D on a screen to access images located within a 3D view. This would necessitate some processing which would probably need to be local to the viewer and this might impose some limits on accessibility due to capabilities of some platforms.
b) Later, we wish to formally evaluate the semantic performance of the summaries. The added variable of a component image's position within a 3D summary would complicate such an evaluation.

6.2 Clustering Method

Clustering based on perceptual data will be used to find suitably representative images for the summarisation.

Clustering is used as a way of discovering or describing structure in multivariate data. There are many methods of clustering, and Everitt (1974) divides the methods into five categories: "Hierarchical" (a classification tree is formed), "Optimisation/Partitioning" (a clustering criterion is chosen and the data are split based on that), "Density"

(concentrations of data form the focus for the clusters), “Clumping” (clusters may overlap) and “Others”. While those methods which discover the structure in data, such as the hierarchical methods, are particularly useful for interrogating the fine structure of data the output in the form of a classification tree is, to a large degree, determined by the data.

We wish to produce a summary in which we have full control over the number of representative images which result in the output. One of the most commonly used methods enabling this (Martinez et al. 2011) is *k-means;* this being one of the partitioning clustering methods. *K*-means allows “*k*”, the number of resulting clusters, to be specified at the outset.

A partially pragmatic decision was taken to proceed with *k*-means as the clustering method to be used. Aside from the ability to specify the number of clusters, the following factors were taken into account: a) *k*-means, being commonly used, is implemented and obtainable in many programming environments, and b) other partitioning methods exist and should the opportunity arise to optimise the clustering, or should *k*-means prove unsatisfactory in some way, alternatives can then be investigated. In short, *k*-means clustering is appropriate and a useful working method.

6.3 Dimensionality Reduction

This section discusses the need for dimensionality reduction to allow the multidimensional perceptual data describing the Abstract 500 image set to inform a set of 2D coordinates for the summaries.

The evaluation of the Abstract500 perceptual data using classical MDS (see 4.6) indicated that the perceptual data described by the similarity matrix was of the order of 12 to 20-dimensional. As the summaries are to be 2D, dimensionality reduction will be needed to represent the perceptual data of the Abstract500. This need for dimensionality reduction would be the case with any image set which might be deployed with the purpose of providing a wide visual vocabulary for feedback.

The choice of dimensionality reduction may influence the effectiveness of the relative placement of the representative images. (Note that the choice of the representative images, being based on clustering in the full 500x500 perceptual similarity matrix, will not be affected by the dimensionality reduction).

In the course of their work on visualisation of multivariate data, Shroeder & Noy (2001) pointed out that establishing the appropriate method of dimensionality reduction for a given purpose was, essentially, a matter of comparing the results achieved with different methods and choosing that which worked best for that given case. In the next subsection four methods of dimensionality reduction are compared in the context of the Abstract500 perceptual data with the purpose of choosing the most suitable method.

6.3.1 Choice of Dimensionality Reduction Method

Four methods of dimensionality reduction were applied to the Abstract500 perceptual data: classical MDS (Cox & Cox, 2001), non-metric MDS (Kruskal, 1964a, 1964b), Isomap (Tenenbaum et al., 2000) and Isomap II (or Landmark-Isomap) (Silva & Tenenbaum 2002). 3D visualisations based on these 4 methods were compared. See Appendix A.

Following the comparison of the four 3D visualisations the following was observed: In general, the distributions of all four were not greatly dissimilar aside from the nonmetric MDS showing the more open distribution. All four showed that the sample themed image groups were clustered as discernable groups (e.g. nature themed images). Image 10, often a singleton during the bootstrap perceptual grouping, was placed away from the other images in the non-metric MDS view, while being more closely embedded amongst other images in the other three views.

Taking into account these observations, non-metric MDS was chosen as the method of dimensionality reduction. This is based on visualisations of its application to the Abstract500 image set. If the dimensionality reduction is to be applied to another image set it may be appropriate to use a different method.

6.4 Rationale for Two-Stage Dimensionality Reduction

3D visualisations based on 3D MDS of the Abstract500 image set are convincing during informal examination. The dimensionality reduction outputs (eigenvalues on classical MDS; stress in non-metric MDS and residual variance in Isomap and Landmark-Isomap (see Appendix A)) show that the first 3 dimensions in the data account for a large proportion of the variation within the data.

It is this convincing nature of the 3D MDS visualisations of the Abstract500 image set that have motivated the decision to use this 3D data as positioning data for the summaries; i.e. the 3D MDS coordinates will inform the 2D position for any representative image when placing it on a summary.

Making use of 3D coordinates from a dimensionality reduction on the whole image set will allow this processing to be done offline and before any feedback image selection and clustering. Leaving a last stage of dimensionality reduction (from 3D to 2D) until after clustering and representative image selection, will allow a 2D representation which better portrays the relationships specifically between only the small subset of *k* representative images. This will allow a more faithful portrayal of the relationships of the representative images relative to each other.

The dimensionality reduction will therefore be applied in two stages:

- Stage One: Reduction from *n*-dimensions to 3D using non-metric MDS, for full image set, offline.
- Stage Two: Reduction from 3D to 2D, after clustering of feedback selection and representative image selection, thus more optimally portraying the relationships between the representative images.

6.5 Method for the Reduction from 3D to 2D

As this may need to be done "on-line" i.e. after the calculation of the representative images for a given summary this should be a low cost process (in terms of processing) to place a low processing load on any application which delivers it.

The input will be the IDs of the representative images and the set of 3D coordinates associated with each. (In the case of the Abstract500 set these will be from 3D non-metric MDS). See Figure 6.2.

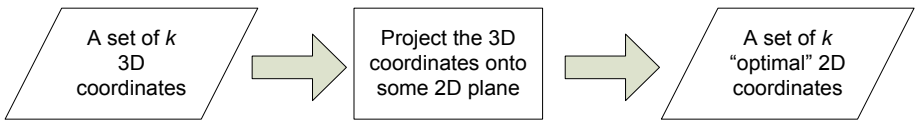


Figure 6.2 - Flow diagram for the final stage reduction from 3D to 2D.

The selection of the plane onto which to project the 3D coordinates was set as being that lying on the triangle between a) the two representative images furthest apart (by

Euclidean distance) and b) the image representing the most popular cluster. This prioritises the preservation of the relationship between the two most distant of the *k* images and that representing the most popular cluster. It uses the two most distant images to define the dissimilarity scope of the visual summary. Projecting the existing *k* 3D points onto a plane defined by three points already within the set of *k* points will be a low cost process (in terms of processing).

Table 6.2 shows pseudocode for this process.

1	**Input**: $C_1..C_k$ ordered by P highest to lowest.
2	Set $Pt_1 = C_1(X_{MDS}, Y_{MDS}, Z_{MDS})$.
3	From $C_2..C_{10}$ find C_x and C_y, the 2 members sharing the largest inter-cluster distance in *3D MDS* space.
4	Set Pt_2 and Pt_3 to be $C_x(X_{MDS}, Y_{MDS}, Z_{MDS})$ and $C_y(X_{MDS}, Y_{MDS}, Z_{MDS})$
5	Set the *Optimal Plane* to lie on triangle [Pt_1, Pt_2, Pt_3].
6	**For** i=1 **to** *k* **do**
7	Set $C_i(X_{OPP}, Y_{OPP})$ to be the orthogonal projection of $C_i(X_{MDS}, Y_{MDS}, Z_{MDS})$ on the *Optimal Plane.*
8	**EndFor**
9	**Output**: $C_1(X_{OPP}, Y_{OPP})..C_k(X_{OPP}, Y_{OPP})$

Table 6.2 - Pseudocode for the final stage of reduction from 3D to 2D coordinates. $C_1..C_k$ is the list of cluster representative images. Each C has a cluster population, P.

Thus, the final reduction from 3D to 2D will be done by projecting the 3D coordinates onto a plane defined by representative image of the most popular cluster and the two representative images furthest apart by Euclidean distance.

6.6 Overlapping Images on Summaries

The works referred to in 5.3 on image summarisation which produced 2D collage-style summaries (Rother et al., 2006), (Lee et al., 2010) and (Egorova et al., 2008) involve processes which allow overlapping images and thus the partial obscuring of some regions of some images in the summaries. This is done for aesthetic reasons. However, as we, later, wish to formally evaluate the semantic performance of the summaries we will ensure no part of an individual image on a summary is obscured, as this could have

an effect on the semantic content or emotive impact of the image and on the results of any evaluation. Thus, the developed summaries will be non-overlapping.

6.7 Method for Rendering the Summaries

An algorithm for rendering the 2D non-overlapping summaries was developed, see Figure 6.3.

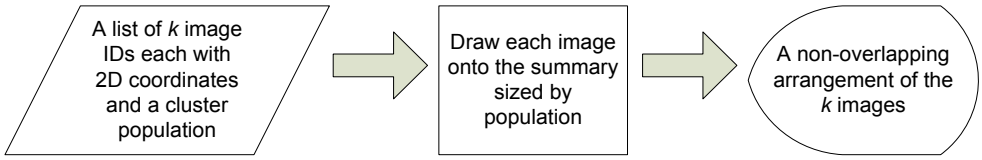


Figure 6.3 - Flow diagram of rendering a visual summary.

Table 6.3 shows pseudocode for this process.

1	**Input**: $C_1..C_k$ ordered by P highest to lowest.
2	Establish the area for the montage based on the device/window and the range of cluster coordinates and populations.
3	Place C_1 on $C_1(X, Y)$
4	**For** $i = 2$ **to** k **do**
5	**If** placement of C_i , sized proportionately, is obstructed
6	Locate alternative placement closest to C_i (X, Y) by heuristic search
7	**EndIf**
8	Place C_i , sized proportionately.
9	**EndFor**

Table 6.3 - Pseudocode for rendering the non-overlapping arrangement of the list of k cluster representative images, $C_1..C_k$. Each C has a cluster population, P and ideal placement 2D coordinates (X, Y), output from the final stage of dimensionality reduction. (See Table 6.2).

Thus with the sorted list and seeking the nearest spot for each successive placement this is a greedy algorithm. i.e. locally optimal but aiming to achieve a globally optimal solution by being so.

6.7.1 The Heuristic Search

An algorithm for the heuristic search of the 2D space on the image summary invoked in step 6 of Table 6.3 was developed and is described below.

When the placement of an image in the summary is obstructed by the edge of the summary or another image alternative placements points are generated producing in effect a search tree. The process is a breadth first search of that tree.

Initially, up to 8 alternate placement points are generated. The angle of displacement of the alternate points can be likened to the points of the compass (north, north-east, east, south-east etc.) while the magnitude of displacement depends on the size and position of the obstruction and the size of the image being placed. The set of new alternate placements is tested and the list of any that are in bounds (within the bounds of the summary space) and not obstructed is compiled. From this list of candidates the alternate point closest (by Euclidean distance) to the ideal point is chosen. Should all of the alternate placements be out-of-bounds or obstructed then a set of alternate placement points for each obstructed alternate placement (but not out of bounds placements) is generated and tested. For search depths one to four only 7 alternate placements are tested, i.e. not that which would vector back to where it originated.

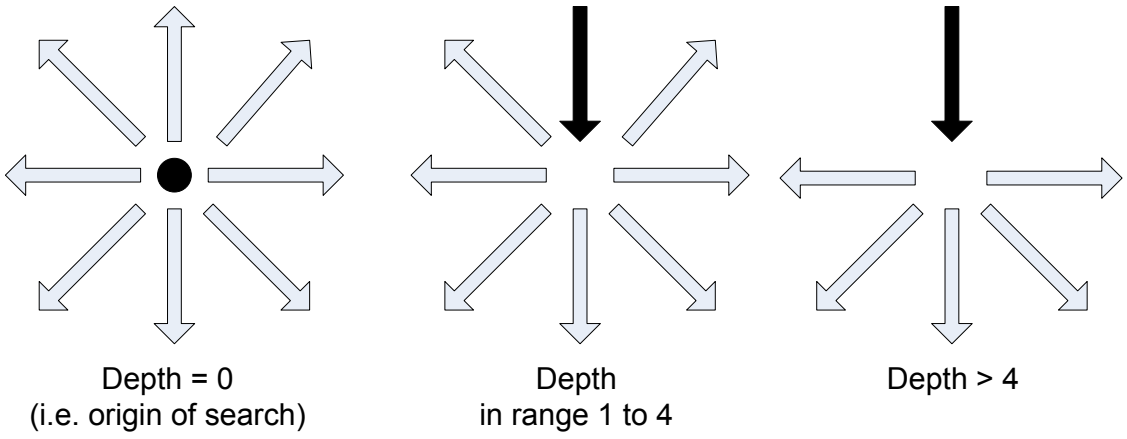


Figure 6.4 - Image summary space search heuristic. There are three search depth cases. Black arrows indicate the incoming search vector. Pale arrows indicate outgoing (or onward) search vectors, their angles of displacement being relative to the incoming vector. The black dot indicates the origin of the search i.e. null incoming search vector.

This creates a search tree which is searched breadth first. The depth of the search is monitored and should it extend beyond four then a modified heuristic is applied to limit

the size of the search (in terms of nodes). This then causes the search to extend further, faster, away from the ideal point, beyond obstructions and resolving any blockage due to congestion which might occur in a corner of the summary. See Figure 6.4.

6.8 Implementation

This section describes the implementation of the steps of the summarisation algorithm.

The summarisation was used during the evaluation described in Chapter 10 therefore refer to that chapter to locate the code for the implementation which is described in this section.

6.8.1 Stage One of Dimensionality Reduction

Stage one of dimensionality reduction was implemented in MATLAB taking, as input, the *Abstract500* similarity matrix and outputting 3D non-metric MDS coordinates (x, y, z,) for each image. This process only needed to be done once as it applied to the whole image set. The list of 3D coordinates was stored as a CSV file.

6.8.2 Clustering

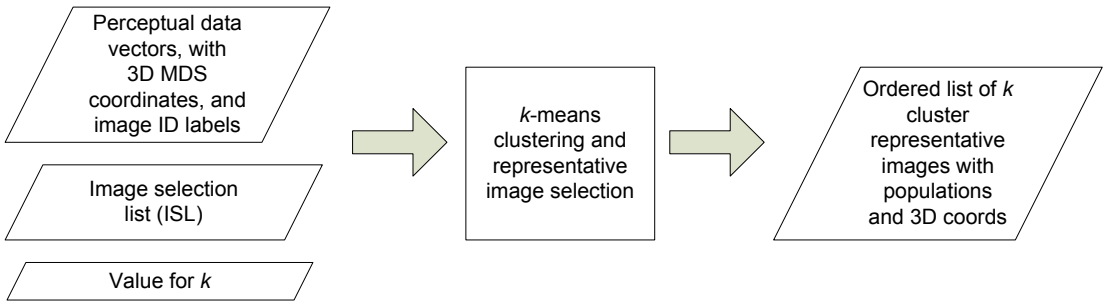


Figure 6.5 - The MATLAB clustering implementation. In the case of the Abstract500 the perceptual data vectors were the rows of the similarity matrix. The representative image for a given cluster was that image from the ISL, nearest, by Euclidean distance, to the cluster centroid.

The clustering was implemented in MATLAB (and the MATLAB Statistics Toolbox). The *k*-means clustering command was invoked, with parameters set to seed the clustering with *k* randomly selected data points. To ensure there were no empty clusters in the output, clustering was repeated until the number of empty clusters was zero. One

of the inputs, an image selection list (ISL), is a list of image IDs representing the *x* images per participant from the *n* participants selecting images from a browser in response to some stimulus. See Figure 6.5. Each ISL was compiled using PHP and MySQL scripts which run queries on the database used to store users' (or experiment participants') image responses.

Setting k for k-means clustering

The factors taken into account in choosing a suitable value of *k* for the *k*-means clustering, i.e. deciding how many representative images to place in the summaries, are set out in the table below.

Factor	As implemented in Chapter 7 Experiment
Display screen resolution	iPads were used to display the summaries. Resolution 1,024x768 pixels with 132 pixels per inch. Each image in the Abstract500 is 128x128 pixels resolution.
Expected number of image selections in each image selection list (ISL) would need to contrast with k for summarisation experiment validity.	60 image selections. (Each ISL may contain multiple instances of any given image due to agreement among participants when representing a term or concept using the images.)
Allowance for diversity within the semantic content of each ISL.	

Table 6.4 – Factors in setting k for k-means clustering. See accompanying text for details.

In practice, for the experiment in Chapter 7, screen resolution was lower priority than image selection list (ISL) size. As the experiment was to test the effectiveness of the summarisation method, there had to be a marked difference between the number of representative images (k) and the number of images selected by each of the experiment participants. During that experiment participants were asked to choose images to represent given terms and the collated image choices or image selection lists (ISLs) summarised using the summarisation algorithm. Each participant was asked respond to each term by choosing three images to represent that term. There were 20 terms. Thus 20 x 3 = 60 images in each ISL. 10 was a pragmatic choice as the value for k taking into account the factors in Table 6.4. There was not time in the experimental schedule to do a more elaborate optimisation. 10 was used as the k value. The summaries, consisting of 10 representative images were evaluated in the experiment described in Chapter 7 and found to be effective compared to the ISL they summarised. It is possible a more effective value for k might be found through experimentation and this may be an avenue for future work.

6.8.3 Stage Two of Dimensionality Reduction

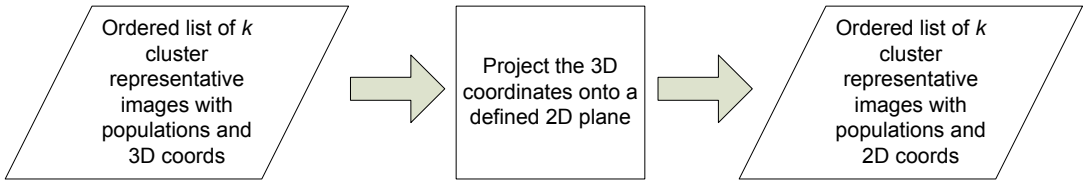


Figure 6.6 - The MATLAB final 3D to 2D reduction implementation. The input list including 3D coordinates is the output from clustering (see Figure 6.5). The algorithm for defining the 2D plane is described in Table 6.2.

The final dimensionality reduction (Figure 6.6) which takes into account the relationships between the k representative images (see Table 6.2) was implemented in MATLAB taking, as input, the ordered list of *k* cluster representative images with populations and 3D non-metric MDS coordinates. The output is the same list but with 2D coordinates rather than 3D. The author's MATLAB code made use of *geom3d* a 3D geometry library (Legland, 2009).

In practice, clustering and stage two dimensionality reduction was combined within one MATLAB script.

6.8.4 Rendering the Summaries

The rendering was implemented in PHP (to load the data) and JavaScript. The JavaScript handled the heuristic search for image placement and the drawing/rendering aspects. Use was made of the Raphael cross-browser JavaScript graphics library (Baranovskiy, 2010) and a collision detection class from a 2D game library (Wallin, 2010). The input is a CSV file containing an ordered list of *k* clusters detailing the following for each cluster: ClusterID, population, representative image ID and ideal image coordinates (X, Y).

6.9 Conclusion

An algorithm, based on *k*-means clustering, was developed to take a list of images selected from an image set with accompanying perceptual data, and summarise it by placing a small number (*k*) of representative images on a two-dimensional, non-overlapping, summary collage. The size (area) of each representative image on the

summary varies proportional to population of the cluster it represents. The algorithm is illustrated in Figure 6.7.

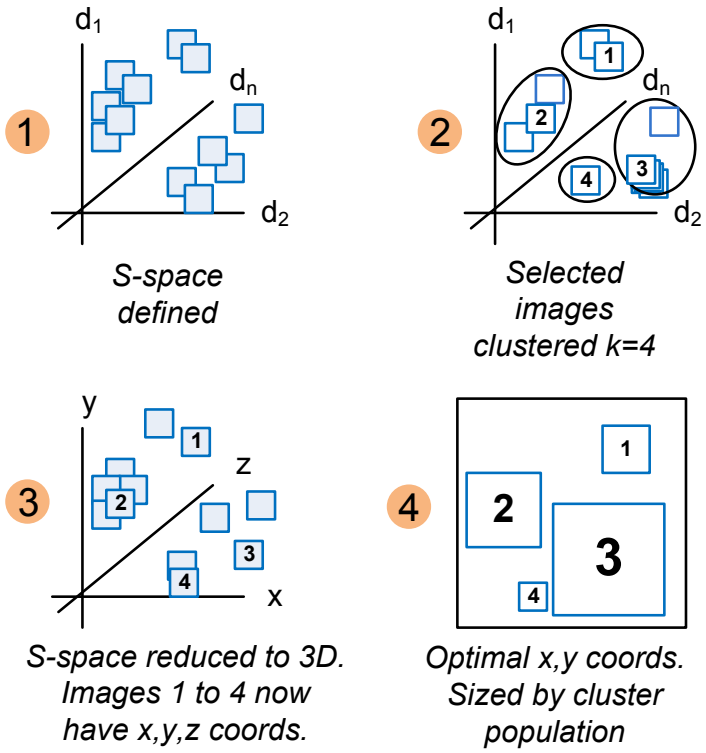


Figure 6.7 - The summarisation method. Step 1: Similarity space (S-space) is defined by the perceptual data for the image set. Step 2: the selected images (or ISL) are clustered in S-space; for simplicity, this example uses k=4; each cluster's representative is that nearest to its centroid. Step3: the first stage of dimensionality reduction down to 3D is done by MDS relative to the entire image set. Step 4: final dimensionality reduction from 3D to 2D is done relative to the k representative images.

The algorithm has been implemented in MATLAB, PHP and JavaScript. The weakness of the implementation is that it does not provide a fully integrated end-to-end web application for the processing of image feedback. The perceptual data and the 3D coordinates from the first stage of dimensionality reduction can be done at the outset before any participant image selections are gathered. However, after the participant image selections are gathered, the clustering requires a MATLAB processing step before the summaries are able to be rendered in a web application.

Despite the above limitation, this implementation of the visual summarisation algorithm will be *sufficient to allow the evaluation of the method* and so satisfies the thesis goal. Expending time to develop the clustering beyond this prototype implementation as an end-to-end web application is outside the scope of the thesis.

For the Abstract500 image set this summarisation can use the same perceptual data as was gathered to inform the construction of the Abstract500 SOM browser.

As a post script to this chapter Figure 6.8 shows an example of a summary produced during Chapter 7 along with screenshots of MDS views illustration of the clustering.

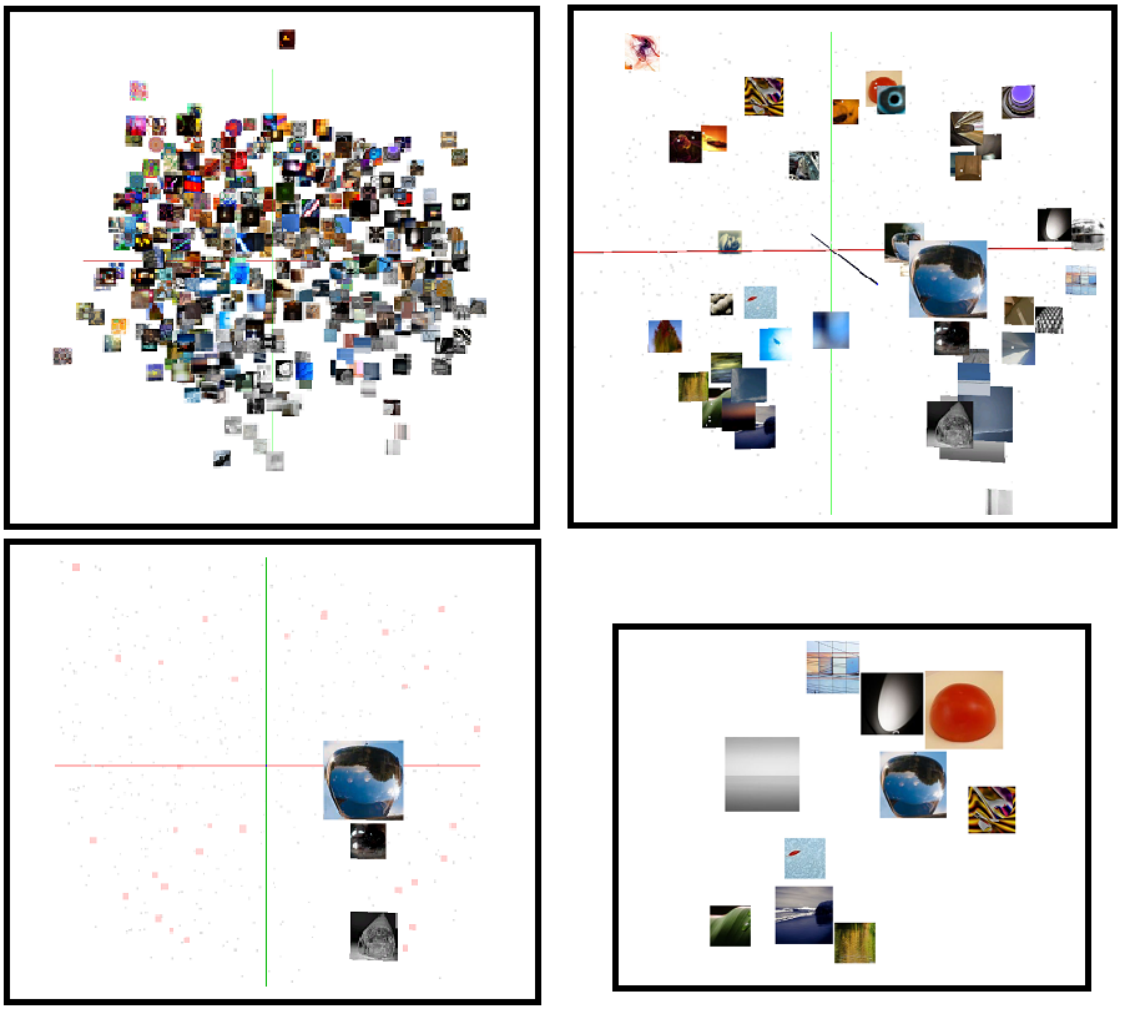


Figure 6.8 - An example summarisation from Chapter 7. ***Top left:*** *The Abstract500 in a 3D non-metric MDS view of S-space;* ***Top right:*** *An image selection chosen by participants to represent "smooth" projected onto the 3D space, sized by popularity.* ***Bottom left:*** *One cluster isolated in the view.* ***Bottom right:*** *The 2D summary. Note the glass orb representing the cluster.*

Chapter 7

Communication Evaluation

One of the goals of this thesis is to "*develop the means to implement the CVFM sufficiently to allow evaluation"*. Chapters 4 and 6 developed two components, the Abstract500 SOM browser and the image summarisation algorithm, as the *means* for implementing the CVFM. This chapter will establish whether these two components are *sufficient* to allow evaluation of the CVFM; i.e. do they work and how well do they work?

Thus, the purpose of this chapter is to evaluate a) the utility of the Abstract500 SOM browser for enabling a crowd to communicate its reaction about an idea to another individual and b) the effectiveness of the summarisation algorithm at producing summaries which communicate what they are meant to have summarised. Both these goals share a common theme, that of *communication evaluation*. With these goals and this common theme in mind, the Communication Evaluation Research Questions (CERQ), set out in Table 7.1, were formulated.

CERQ No	Communication Evaluation Research Question
1	To what degree can meaning be communicated by the image selections of a crowd from the Abstract500 to another individual?
2	Are the visual summaries of image selections, created using the summarisation algorithm, more or less effective at communicating meaning than the image selections which they summarise?

Table 7.1 - The Communication Evaluation Research Questions (CERQs).

A single communication experiment was devised to address these research questions. However, as it addresses both Communication Evaluation Research Questions 1 and 2 the experiment is introduced in two parts (or aspects) diagrammatically, using the ideas and icons from the CVFM diagram in the Thesis Introduction (Figure 1.1). The first aspect addresses CERQ No.1.

Experiment Aspect 1: Communication (addressing CERQ No.1)

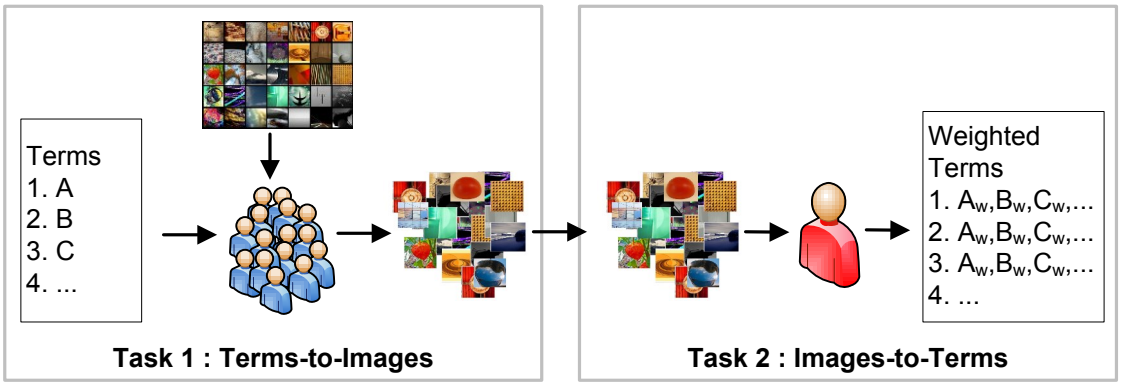


Figure 7.1 - Experiment Aspect 1: Communication- addressing CERQ.1.

Aspect 1 (Figure 7.1) addresses CERQ No.1 using two tasks for human participants. In Task 1 participants, representing the crowd, view terms (one at a time) as stimuli and select images to represent those terms. The image selections for each term, or *term image selections* (*TIS*), are collected. In Task 2 the TIS are shown as stimuli to a different participant group (representing a designer) unaware of the intended meaning of each TIS. For each TIS the participants output the full set of terms, assigning each a weighting according to their judgment of the degree to which the meaning of each term is present in that TIS. **The output weightings for each term are used as a metric for the effectiveness of communication for each TIS**; e.g. if Task 2 participants viewing the TIS for term *A* tend to allocate a high weighting for term *A* in that stimulus relative to their weightings for the other terms, then communication of term *A* using the Abstract500 SOM browser will be judged successful. The success of the communication of each term relative to other terms can be used to determine strengths and weaknesses of the Abstract500 SOM browser for communication.

Experiment Aspect 2: Comparison of Communication (addressing CERQ No.2)

Aspect 2 addresses CERQ No.2 by, *in addition to Aspect 1*, generating summaries from the term image selections (*TIS*) and having the human participants in Task 2 judge the meaning content of the summaries in the same way as for the TIS. The output of Aspect 2 of the experiment is obtained when the **effectiveness of communication for each TIS is compared with that for their corresponding summaries.** (Figure 7.2).

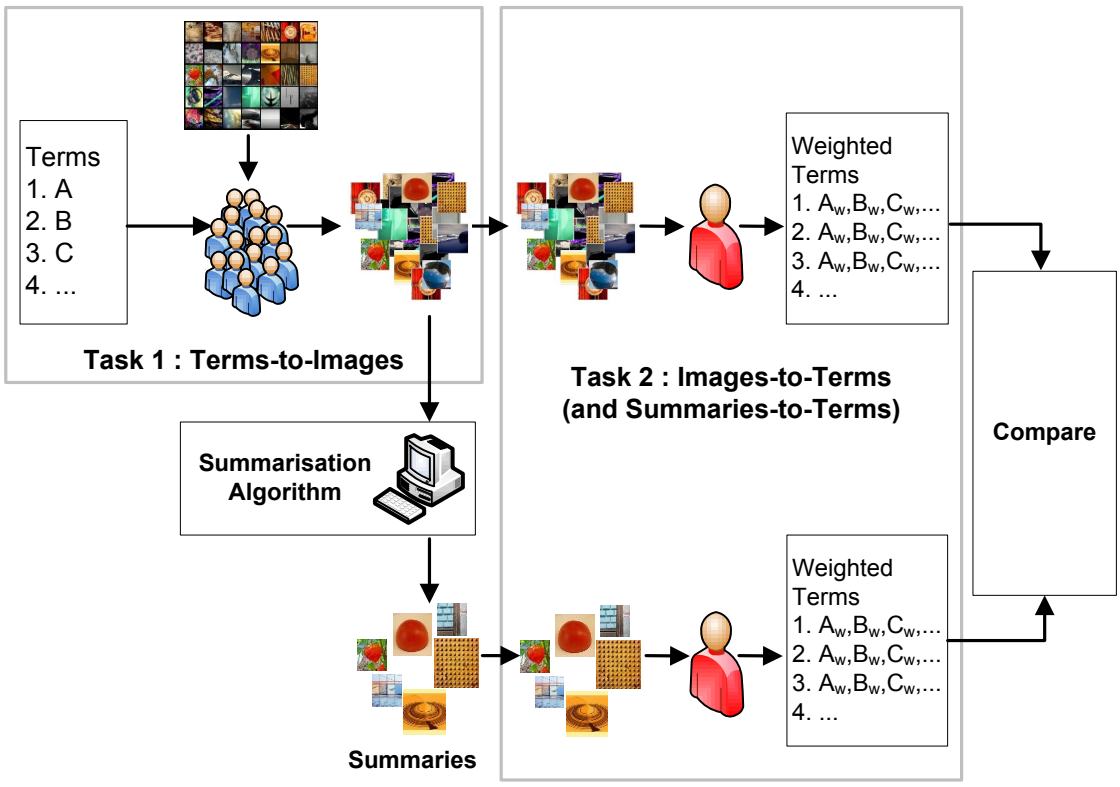


Figure 7.2 - Experiment Aspect 2: Comparison of communication of summaries with image selections- addressing CERQ No.2.

Having set out this overview of the Communication Evaluation Research Questions and Experiment, the remainder of this chapter describes the details of the experiment and its results. Section 7.1, introduces what experiment participants will do, points out that the experiment is really both an observational study *and* an experiment, details the methodology and the experimental variables, sets out the specifics about how participant observations will be made, describes the terms to serve as stimuli, and lastly relates the details of participant recruitment and task conditions. Sections 7.2 to 7.5 set out how the Task 1 observations were gathered to form the lists of image selections, the generation of the summaries and the visualisation of these outputs from Task 1. The Task 2 interface and recording method are described in 7.6. Section 7.7 sets out the results from Task 2 and the comparison of the performance of the summaries vs. the image selections. Finally, Section 7.8 concludes the chapter by revisiting the Communication Evaluation Research Questions to establish the answers and proposes the next steps in the thesis in the light of the conclusions of this chapter.

Appendix E is the appendix associated with this chapter.

Published work

The experiment in this chapter features in published work: Kalkreuter et al. (2013).

7.1 Experiment Design

The communication experiment would consist of two tasks as depicted in (Figure 7.2). In Task 1 the participants, acting as the crowd, stimulated by a number of terms, would output a number of term image selections (TIS) corresponding to those terms. Prior to Task 2 each of the TIS would be fed to the summarisation algorithm as input producing associated summaries as output. Together the TIS and associated summaries would comprise the stimuli for the Task 2 participants and these stimuli would all have some intended meaning from Task 1. The Task 2 participants would be naïve of the intended meanings for the stimuli, but would be shown all the terms and asked to weight them (or rate them) according to their judgment of the degree to which the meaning of each term is present in a given stimulus (a TIS or a summary).

7.1.1 Both an Experiment and a Study

Although the communication experiment is described above as "an experiment", it would be more accurate to describe it as containing both an observational study and an experiment.

Observational Study for CERQ1

The results from the Task 2 weighting (or rating**)** of meaning content of the Task 1 term image selections (TIS) and summaries would address CERQ1, "*To what degree can meaning be communicated by the image selections of a crowd from the Abstract500 to another individual?".* Strictly speaking this is an observational study; we would serve word stimuli in Task 1 and form the visual stimuli for later use in Task 2. Then we would observe how the Task 2 participants rate those visual stimuli.

Experiment for CERQ2

The ***comparison*** of meaning content (or term ratings) of the TIS and their corresponding summaries after Task 2, would constitute the experimental manipulation of a variable. The variable would be the type (or *format*) of stimulus (image set or summary) for a given term. This would address CERQ2, "*Are the visual summaries ...*

more or less effective at communicating meaning than the image selections which they summarise?"

7.1.2 Methodology

A quantitative method would be used as it was possible, indeed expected, that this would lead to a clear answer at least for CERQ2. Participants in Task 1 would view terms (verbal stimuli) and choose images to represent those terms. This would produce image selections and after processing, summaries, all with ***intended meanings***. These image selections and, summaries would become visual stimuli for Task 2. In Task 2 participants would view these visual stimuli, each having an intended meaning, and report the degree to which the meaning of all of the terms (including the intended meanings) was present in the stimuli using visual analogue scale (VAS) items (Reips & Funke, 2008) (Hofmans & Theuns, 2008). Thus a Task 2 participant's observation of a single visual stimulus would produce a set of numbers (interval data) representing their VAS ratings for that stimulus. In 7.1.3 "Variables", the use of these numerical data will be described.

A repeated measures paradigm would be used. All participants would view all stimuli. (This was later modified with each participant seeing a random selection of half the stimuli balanced so they saw an equal number of image selections and summaries; see below).

7.1.3 Variables

Independent variable

The independent variable was to be ***the format of the visual stimulus*.** This would have 2 conditions:

1. Image selection list
2. Summary

Dependent variable

There would be one measurement:

- The ***relative meaning content*** of all the feedback terms in each stimulus as rated by the Task 2 participants.

Frequency of first rank for intended meaning (f-1st)

The measurement of the single dependent variable, *relative meaning content,* would involve the subsidiary measurement, for each visual stimulus, of each participant's report (or rating) of the degree to which the meaning of each of the feedback terms is present in the stimulus. The terms for that stimulus could then be ranked on this rating revealing the particular participant's top rated term for that stimulus. If that participant's top rated term for that stimulus was the ***intended meaning*** then this would become an occurrence of ***first rank for intended meaning*** for that stimulus. The frequency with which this occurred, ***f-1st***, normalised for the number of participants, would be a metric for the communicative effectiveness of that stimulus. (Standard competition ranking would be used; i.e. a score's rank is always one plus the number of greater scores. This means a rating which ties for first place counts as first rank.)

The experimental result for CERQ2 (effectiveness of summarisation) would be obtained by a correlation analysis comparing the ***frequency of first rank for intended meaning*** (***f-1st***) for image selection lists with that for the corresponding summaries. The observational study results for CERQ1 (effectiveness of the Abstract500 browser for communication) would compare the ***f-1st*** for all the stimuli and the ***f-1st*** that would be expected had Task 2 participants rated the meaning content randomly, to gauge relative effectiveness of communication across the feedback terms.

7.1.4 Visual Analogue Scale (VAS) Item Wording

The wording devised for the VAS items in Task 2 is shown in Table 7.2. There would be one VAS item per feedback term for each stimulus.

VAS Item Wording	Anchor1	Anchor2
Measure: Meaning content		
Is the meaning of the word or phrase present in the pictures?	Clearly Present	Clearly Absent

Table 7.2 - Pilot VAS item wordings.

7.1.5 The 20 Feedback Terms

As the domain of fashion design was one of the original inspirations for the CVFM a sample of terms descriptive of material properties would be appropriate for that domain and to serve as an abstraction for all material properties. The importance of emotions in design was established earlier in this thesis. Thus a sample of emotive terms would serve as an abstraction of all emotion terms.

Thus, the set of terms selected to be used as stimuli in Task 1 and to assess meaning content in Task 2 consisted of

a) 10 terms descriptive of material properties (e.g. flexible and textured) selected from terms output by a study which asked naïve participants to volunteer words describing fabrics (Methven et al., 2011), and
b) 10 emotive terms (e.g. "astonishment, surprise" and "disgust, repulsion") selected from an emotion model (Scherer, 2005).

The 10 terms descriptive of material properties

Methven et al. (2011) sourced 78 words used to describe fabrics from technical journals and from naïve participants. The perceived similarity between the terms was defined having participants free group them based on their meanings. This similarity data was visualised using a dendrogram. Methven et al. exposed 11 clusters (by cutting the dendrogram at a particular height). Two of the clusters contained terms such as "natural" and "even" and also "hot" and "cold", which were less relevant to fabric material than the other clusters. Thus, for this purpose, these two clusters were set aside. One term was selected to represent each of the remaining nine clusters. Additionally one further term from the largest cluster was selected to give 10 terms in total.

The 10 emotion terms

The Geneva Emotion Wheel (GEW) model of emotions, being a model used often in research referring to emotion (e.g. Siegert, et al. (2011), Pammi, & Schroder (2009) and Soleymani & Pantic (2012)) was selected as a source for emotion terms. Version 2 of the model as shown in Sacharin et al. (2012) consists of 20 emotion terms arranged symmetrically around the two dimensions of valence (positive/negative) and control (sometimes termed arousal). Five terms from the negative valence and five terms from

the positive valence regions of the wheel were chosen thus offering a balanced set of 10 positive and negative emotion terms from the wheel model.

As stated already in the Publications section under heading “Padilla and the work of Chapter 7” (p iii). The researching and choice of these terms was done by Padilla (2011) when creating an application for prompting participants to choose images (from a different image set) to represent terms; that application being also used for Task 1 in this thesis as stated in 7.2.1).

The terms are listed in Appendix E p.228.

7.1.6 Participant Recruitment and Task Conditions

Participants were to be sought from the university campuses and a gender balance would be aimed for. To give the task broad appeal and so attract as wide a range of participants as possible the tasks were to be done on iPad tablet computers, to be of a relatively short duration, and be portable to avoid participants having to organise appointments and travel to the lab. The compensation to be offered would be 100g of chocolate or snack of similar value for Task 1 (about 20 minute’s duration). The duration of Task 2 was around 30 to 40 minutes and so the compensation offered was £10 in Amazon vouchers.

7.2 Task 1 - Terms-to-Images

7.2.1 Interface and Recording Method

It should be noted that, when the author joined the project that is the subject of this thesis, some practical work had already been done by Padilla to investigate the possibility of communicating ideas with images using a different image set. That work, consisting of an application to prompt participants to choose images to represent a chosen set of terms, was passed to the author in a private communication (Padilla, 2011). The availability of that application, the chosen set of terms, and their suitability for the evaluation influenced the design of the experiment. However, using them saved the cost of developing a new application for the purpose.

The experiment application served the 20 stimuli terms in a random order prompting the participant to choose 3 different images from an image set in SOM browser form. The application was adapted to present the Abstract500 SOM browser as the interface for image selection. It recorded the participant's image selections in a database. Details of the Task 1 application can be found in Appendix E p.228.

7.2.2 Work Flow

Figure 7.3 shows the workflow for Task 1. The terms would be presented in a random order to participants. A response would require three images so as not to restrict the participant to one region of the image set.

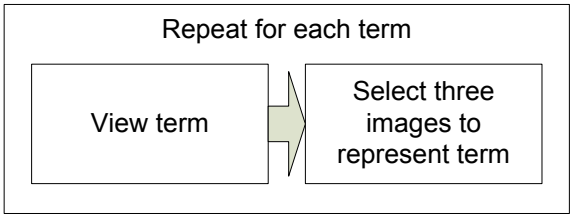


Figure 7.3 - Workflow for Task 1.

7.3 Task 1 - Terms-to-Images Results

7.3.1 The Conduct of the Tasks

20 participants (10 male) were approached in various areas on the campuses and invited to take part. The task was explained. They were shown a demonstration of how to select images from the browser. They were handed an iPad ready to start and left to do the task. A sheet of dictionary definitions was provided in case any participant was in doubt about the meanings. The administrator (the author) withdrew but remained nearby during the task to provide support. Task progress could be monitored remotely on a laptop to ensure smooth progress. Mean time on task excluding one outlier was 25 minutes (median: 25; SD: 5.0; max.: 32; min.: 15). The outlier participant took 72 minutes and found the image browser particularly fascinating. A consistent iPad set-up (e.g. brightness) was used to minimise variation in image presentation.

Figure 7.4 - Participants undertaking Task 1 using iPads.

7.3.2 The Data

Using database queries, the image selections were assembled into CSV lists for each of the 20 terms. Each list of image selections contained 60 image IDs (three per participant). These image selection lists became the input to producing the summaries.

7.4 Producing the Summaries

The image selections were processed using the summarisation algorithm. In practice this involved a MATLAB script which produced 20 visual summary definition CSV files. These definition files were then used by the summary rendering web application to display the summaries. See Appendix E p.228.

7.5 Viewing the Task 1 and Summarisation Output

A web application was created which allowed the image selections and summaries for all 20 feedback terms to be viewed. See Appendix E p.228. Subjectively, the different terms all seemed to have stimulated different image selections with many containing repeated images which indicated that on some terms there may be some agreement on images for terms among participants. The summarisation was functioning. It remained to be seen how the summaries would perform semantically.

7.6 Task 2 - Images-to-Terms

7.6.1 Work Flow

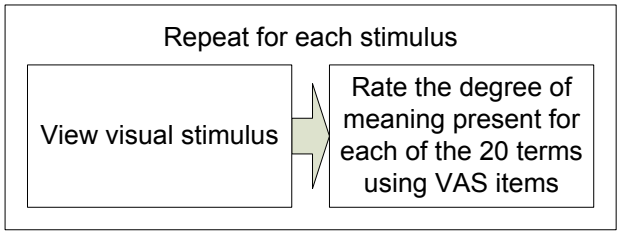


Figure 7.5 - Workflow for Task 2.

Figure 7.5 shows the workflow for Task 2. The visual stimuli (image selections and summaries) would be presented in a random order to participants. There were 40 visual stimuli in all, 20 image selections and 20 corresponding summaries. It was realised that requiring a participant to rate 20 meanings for 40 stimuli (a total of 800 judgements) would make the task too long. Thus each participant would be served a random selection of half the stimuli.

7.6.2 Interface and Recording Method

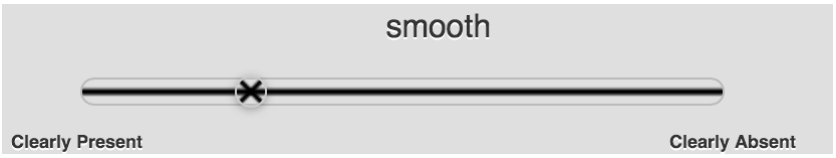


Figure 7.6 - One of the 20 VAS items to be set for each stimulus. In addition participants viewed (and could recall at any time) a dialogue containing this question: "Is the meaning of the word or phrase present in the pictures?". The first tap on the VAS scale caused a draggable cross to appear.

An interface was developed. It served the stimuli and recorded participant VAS item ratings in a database. The stimuli were served according to stimuli packets, generated and stored ready in a database. See Figure 7.6. (Details of stimuli packet compilation and the application interface can be found in Appendix E p.229.) The VAS readings ranged from 0 to 319, based on the number of pixels used to display the scale in the interface application calibrated before the experiment. (Reips & Funke, 2008).

The application would be run on two iPads simultaneously (over wifi Internet) for each participant. One would display the stimuli and the other would enable VAS ratings and control progress. This was achieved simply in the application by a) the slave display iPad running a part of the application which frequently polled a database field to check what stimulus it should display, while b) the master iPad ran a part of the application which altered the value in the polled field when the participant tapped the "next stimulus" button. Participants were briefed to check that the stimulus changed and that a stimulus number indicator on both iPads matched before proceeding.

Figure 7.7 - Two iPads, master and slave, during Task 2. The master (left) recorded ratings and controlled progress. The slave (right) displayed stimuli, in this case an image selection list.

Image summary stimuli fitted comfortably on the iPad display. The image list stimuli were a tight fit. Each of the component images was displayed at 107 x 107 pixel resolution. This was as close as possible to the 128 x 128 resolution (84%) at which they were presented in the Task 1 SOM, while still having the full image selection displayed without scrolling.

To avoid experimental bias due to VAS item positioning within the master display and scale anchor position (left or right), the order of presentation of these was randomised. See Appendix E p.229 for how this was done.

7.7 Task 2 - Images-to-Terms Results

7.7.1 The Conduct of the Tasks

60 participants (30 male) were approached in various areas on the campuses and invited to take part. The task was explained. The author set up the two iPads by logging them in sequentially as master and slave using unique trial login codes. They were handed to the participant who was left to do the task. The administrator (the Author) withdrew but remained nearby during the task to provide support. Task progress could be monitored remotely on a laptop to ensure smooth progress. Mean time on task was 33.5 minutes (median: 30; SD: 10.5; max.: 61; min.: 16). An iPad set-up checklist (e.g. brightness) was followed by the experiment administrator before each session to ensure uniform stimulus display.

7.7.2 The Data

The data from the VAS scale items was gathered by running queries on the recording database. For each participant there were 400 VAS ratings (one per term for 20 terms for 20 stimuli). For each stimulus there were 30 sets of VAS ratings (one per participant viewing each stimulus); each set consisted of 20 VAS ratings (one for each term). The VAS ratings were interval data consisting of integer values.

7.7.3 Frequency of First Rank for Intended Meaning (*f*-1st)

To recap, *f*-1st (fully described in 7.1.3) is the frequency with which participants ranked a visual stimulus' intended meaning first among 20 meanings (or terms). Analysis of the VAS ratings of all the terms for all the stimuli revealed the *f*-1st figures which were normalised (0 to 1). (Detailed figures are in Appendix E p.230.) These are shown in Figure 7.8 along with the frequency level that would be expected had the participants rated the terms randomly for all stimuli. (This was established by generating random simulated studies (Kalos & Whitlock, 2009). (The probability of any particular term from 20 terms being ranked first for a stimulus was calculated to be 0.0515 over 500 random simulated studies consisting of 1000 observations each. Note it is not 0.05, 1/20, due to the probability of there being a tie for first rank.)

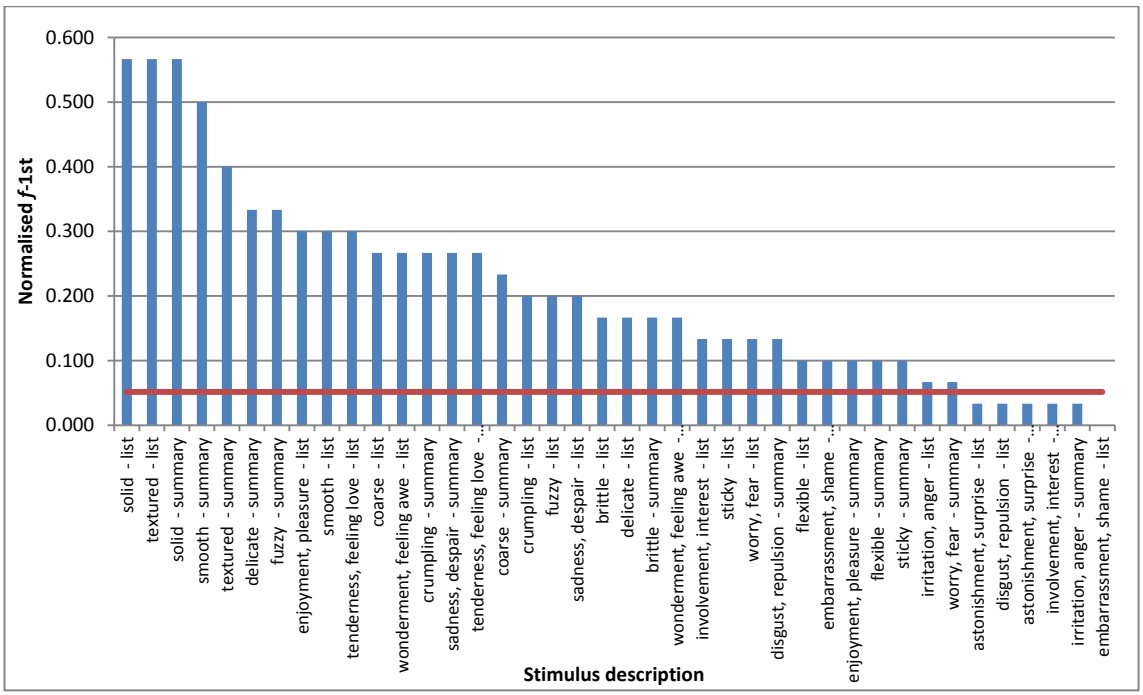


Figure 7.8 - Bar chart showing normalised f-1st for the 40 stimuli with the expected random level shown as a red horizontal line.

Figure 7.8 shows that, for example, half of the participants who viewed the *smooth summary* stimulus rated the term, *smooth*, as their top ranked term for that stimulus (The *smooth summary* stimulus being the summary constructed from those images chosen by the Task 1 participants to represent, *smooth*).

Thus, some of the stimuli conveyed their intended meaning at several times the random probability level, while others performed at or close to the random level. This is evidence that the Abstract500 image browser has varying effectiveness for communicating terms. However, it is also evidence that some communication did take place. The figures were analysed further below.

7.7.4 Comparing Communication of Descriptive Terms and Emotive Terms

The *f*-1st for the stimuli whose intended meanings were *descriptive*, were compared with those for *emotive* stimuli by comparing the means of the *f*-1st figures for those groups of stimuli (Figure 7.9).

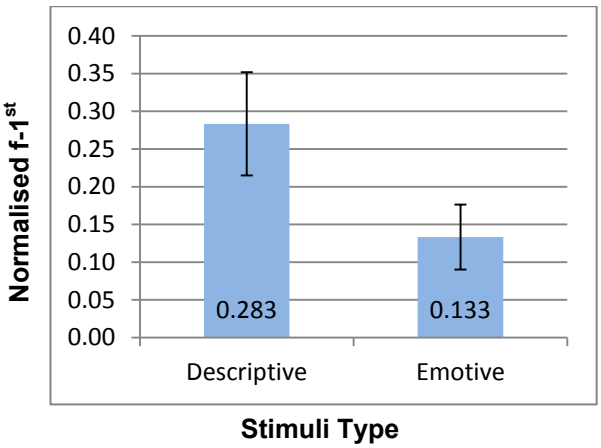


Figure 7.9 - Mean normalised f-1st, descriptive vs. emotive stimuli. Error bars show 95% confidence limits (N=20; 40 stimuli in total).

An independent *t*-test was used. (The hypothesis being that the two means are different and the null hypothesis being that the means are the same). It showed that the mean *f*-1st for stimuli representing descriptive terms (M=0.283, SE=0.036) was significantly greater than for emotive terms (M= 0.133, SE=0.023), $t(38) = 3.543$, $p< 0.05$. This represents a large effect (Field 2009), $r = 0.498$. (Both distributions were tested for normality and passed. See Appendix E p.230.)

This comparison shows that the Abstract500 image set was more effective for communicating descriptive terms than for emotive terms.

7.7.5 Comparing Communication of Summaries and Image Selection Lists

The *f*-1st for the image selection list stimuli (*lists*), were compared with that for the summary stimuli in two ways: by comparison of means and a correlation analysis.

Comparison of means

Figure 7.10 illustrates the two means. A repeated measures *t*-test was used. (The hypothesis being that the two means are different and the null hypothesis being that the means are the same.) It showed that the mean *f*-1st for image list stimuli (M=0.207, SE=0.034) was not statistically significantly different $t(19)=-0.141$, $p>0.05$, to that for summary stimuli (M=0.210, SE=0.035). The *p*-value is greater than 0.05 and not significant at the 95% confidence level with an effect value of $r=0.033$ ($p=0.89$). (Both distributions were tested for normality and passed. See Appendix E p.230.)

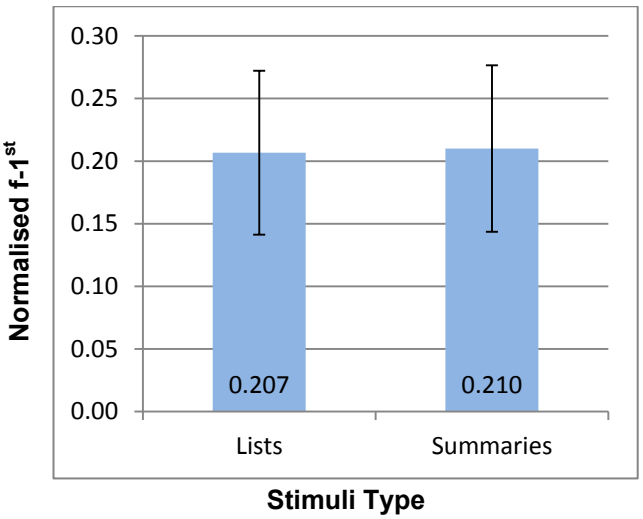


Figure 7.10 - Mean normalised f-1st, for lists vs. summaries. Error bars show 95% confidence limits (N=20; 40 stimuli in total).

Correlation analysis

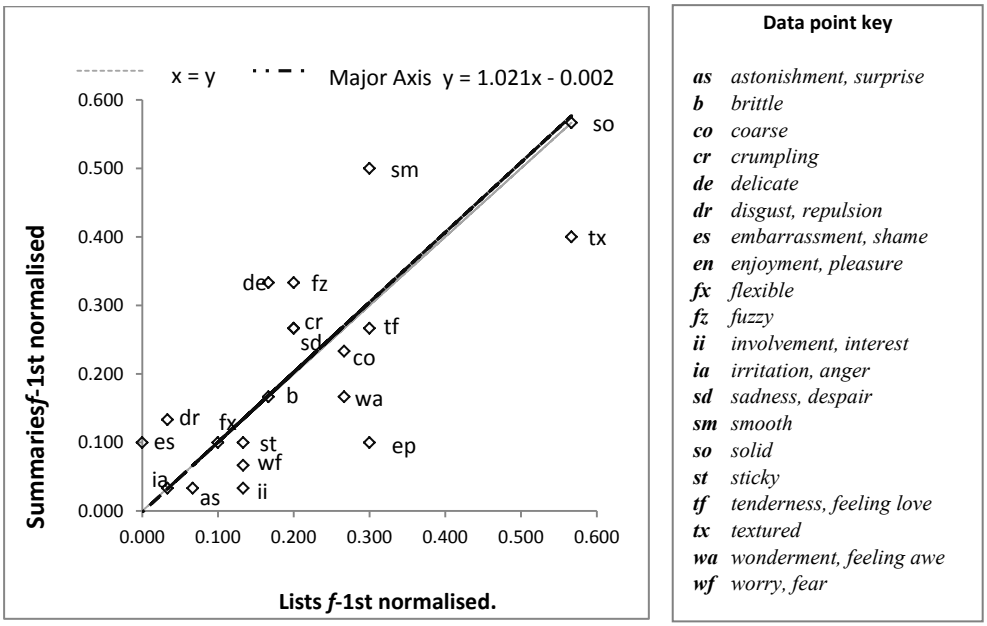


Figure 7.11 - Scatter plot: normalised f-1st lists vs. summaries. The thick broken line is a major axis regression line of best-fit for both x and y. The thin line represents where an ideal one-to-one correlation would lie. The two lines lie very nearly one on top of the other.

The correlation analysis has two components, a *Pearson correlation analysis* and a *Major Axis regression*. A major axis regression calculates a line of best fit for both *x* and *y* components in two distributions (Swan & Sandilands, 1995). The Pearson Correlation Coefficient (PCC) calculation for the two distributions revealed that r = 0.77. This shows that the two distributions are strongly correlated (Field, 2009). Figure 7.11 shows a scatter plot of the *f*-1st for image selection lists vs. their corresponding

summaries. It also shows the major axis ($y = 1.021x - 0.002$) which lies almost coincident with an ideal one-to-one correlation of $y = x$.

Thus the correlation analysis, consisting of a) a strong correlation by PCC ($r = 0.77$), and b) a major axis regression line-of-best-fit for the distributions' x and y components being almost coincident with a perfect one-to-one correlation, is strong evidence that summaries of the image selection lists are as effective at communicating meaning as the image selection lists from which they are generated.

7.8 Conclusion to Chapter 7

In this section, in 7.8.1, the communication evaluation research questions, established at the start of the chapter, are revisited in the light of the results and then in 7.8.2 next steps to address the exposed area of weakness in the Abstract500 image set are presaged.

7.8.1 The Research Questions Revisited

At the beginning of the chapter two communication evaluation research questions were set out. These are revisited in Table 7.3 with answers in brief. Following the table are the answers in full.

CERQ No	***Communication Evaluation Research Question* with Answers**
1	*Can meaning be captured by selections of a crowd from the Abstract500?*
	Yes, but with varying effectiveness. Descriptive meanings were captured significantly better than emotive meanings. It may be effective enough for subjective or impressionistic communication.
2	*Are the visual summaries of image selections, generated using the summarisation algorithm, more or less effective at communicating meaning than the image selections which they summarise?*
	The summaries communicate as effectively as their image selection lists. The summarisation algorithm works for the Abstract500 set

Table 7.3 - Communication evaluation research questions.

CERQ1 has been addressed in 7.7.3 and 7.7.4 by examining the communication performance of the Abstract500 SOM browser a) for the 20 different individual terms and b) the two groups of terms, descriptive and emotive. The effectiveness of the communication varied with the term but was several times random performance level

for some terms. Communication was significantly better for descriptive terms compared to emotive terms. Even the best performance (e.g. solid summary, f-1st of 0.57) would not be described as constituting an unambiguous and precise form of communication. However, it may be enough to allow crowd users to convey their impression of a design idea in a subjective or impressionistic way.

CERQ2 has been addressed in 7.7.5 by the comparison of the communicative effectiveness of the image selection lists with the summaries generated from those lists. The strong correlation of the communication performance of the lists with their summaries shows that the summarisation algorithm, with *k* set to 10, works in the context of the Abstract500 image set.

7.8.2 Next Steps

While image selections from the Abstract500 performed much better than the random performance level for many of the terms, communication was significantly better for descriptive terms on average than for emotive terms. The poor performance of the Abstract500 for communicating emotion terms has exposed that as a weakness of the image set. Relying on the Abstract500 alone when evaluating the CVFM, would risk failure if it is hoped to communicate emotion. Thus, a further image set to support emotion communication should be sought.

Chapter 8

Constructing the Emotive SOM Browser

One of the conclusions of Chapter 7 was that a further image set and browser offering images suited to communicating emotions was required as part of the means to enable evaluation of the CVFM. Thus, this chapter sets out to procure an image set the requirements for which are set out in Table 8.1.

PISR No	Primary Image Set Requirement (PISR)
1	The set must a) be communicative of emotions b) those emotions should be suitable for design feedback; and c) if possible an even spread of emotions should be sought to reduce the risk of biasing the feedback
2	Data must exist (or be obtainable) on each image, suitable to allow a) deployment of the image set in a SOM browser, thus permitting user interaction similar to the Abstract500 in the SOM browser; b) summarisation of selections from the set; and c) a degree of control over the emotion content in the set (to help with 1 b) and c) above).
3	The images must be free to use.
4	There must be enough images in the set to offer users a wide choice and enough such that selections drawn from the set merit summarisation.

Table 8.1 - Requirements for an emotive image set to be used along with the Abstract500 to enable evaluation of the CVFM.

This chapter describes the investigative and practical phases, which included investigating existing image sets, but which culminated in the building of a new emotive image set, *Emotive204*, meeting the requirements.

Early in the investigation phase it was realised that, because emotive images are highly varied in content, similarity data based on free grouping such as that obtained for the Abstract500, would be likely to contain dimensions due to colour, object features extraneous to emotion content, and, other irrelevant dimensions, diluting and even confounding any emotion dimensions with non-emotion noise. In addition it was realised that a full spectrum of emotions a) would not be required and b) was actually undesirable for design feedback (see 8.4.1). Thus it was decided that, to meet both requirements 1 *and* 2, a form of emotion categorisation would be required a) on which

to base the data to inform deployment in the SOM browser and b) to allow some control over the emotion content of the set. See Table 8.2.

SISR No	Secondary Image Set Requirement (SISR)
2.1	The image data should include a form of emotion categorisation.

Table 8.2 - Secondary requirement for the emotion image set.

The rest of the sections in this chapter form two groups, investigation and practical:

Investigation sections: The first three sections (8.1 to 8.3) briefly investigate the background to emotions and images, existing emotion image sets and models of emotion.

Practical sections: The remaining sections describe the practical steps in building a new emotive image set and browser. Section 8.4 details the gathering of 2000 candidate images. Section 8.5 sets out how emotion category data making up an emotion profile for each image was collected thus forming the *Emotive2000*. Section 8.6 describes how the emotion profiles were used to filter the image set and to assemble a balanced portion of it in a SOM browser, the *Emotive204*, ready for use in evaluation studies. The penultimate section (Section 8.7) the *Emotive204* assembled into a SOM browser based on a specific aspect of the emotion profile data for that subset of images.

Finally, in Section 8.8 the image set requirements are revisited and the outputs of the chapter are summarised.

Appendix D is the appendix associated with this chapter.

Published Work

The image set developed in this chapter, along with its accompanying emotion profiles, feature in published work: Robb et al. (2015a) and Robb et al. (2015b).

8.1 Emotion and Images

In this section and its subsections first we discuss the use of images in work in psychology on emotions pointing out one specific emotion image set. Then we show that work has been done to categorise images by emotion. Finally we discuss the complicating factors of image semantics and the psychophysical properties of images.

Image databases have been used in the study of emotions by a number of researchers (Keil et al., 2002) (Meagher et al., 2001) (Hariri, 2003) (Delplanque, 2007). Indeed one image database, The International Affective Picture System (IAPS) (Lang & Bradley, 2007) was developed for the specific purpose of inducing emotional responses in experimental subjects and enabling such emotion studies. It includes images intended to provoke emotions ranging from the highly arousing, such as those depicting erotic subjects, to the most negative, such as those depicting body mutilations. Thus it seems likely that images can be found to be used for our intended design emotion communication.

8.1.1 Emotion Categories for Images

The IAPS images were characterised by three emotion dimensions including, for example, valence which quantifies the degree to which an emotion is positive or negative; e.g. a smiling face would be highly positive valence and a dead body would be low negative valence (see 8.3.1). However, Mikels et al. (2005) categorised some of the IAPS images (into categories such as *fear* and *sadness*) showing a) that emotion categorisation for images is possible and b) that images can have more than one emotion category.

Thus, the requirement for an emotion categorisation, SISDR 1, should be able to be met.

8.1.2 Image Semantics and Visual Properties

Other work on images and emotion has shown that the emotion affect of an image can be as a result not only of the semantic content but also of the visual properties of the image itself. Delplanque et al. (2007) showed that spatial frequencies in an image have effect on emotion. Spatial frequencies within an image are aspects such as sharp defined edges or contrasts (high spatial frequency) and orientation or shape proportion (low spatial frequency). This added complication means that an image within a context of some presentation (e.g. a web page) might have a semantic context, perhaps approximated by the text in the web page directly captioning an image, but due to the image properties, perhaps due to some artefact of the photographic process, it may have a conflicting or additional emotion affect aside from the semantic one. Also the physical size of an image can affect these spatial frequency properties and the emotion affect due to them.

Thus two aspects of emotive imagery are indicated here which may impinge on our aim of creating a browser for emotion communication and on the performance of any browser we do create, a) the context in which we find an image, may not adequately describe its emotion affect, and b) if we alter an image's size we may change its emotion affect.

8.2 Existing Emotive Image Sets

There are image sets established for the study of emotions by psychologists. Perhaps the best known is the International Affective Image System (IAPS) (Lang & Bradley, 2007), a set of images for which there are mean ratings with standard deviations for the dimensions of valance, dominance, and arousal. There is also some categorical data on a subset of IAPS (Mikels, 2005). Other image sets have been established since IAPS e.g. the Geneva Affective Picture Database (GAPED) (Dan-Glauser & Scherer, 2011), and the Necki Affective Picture System (NAPS) (Marchewka et al., 2014). These image sets have all been set up to facilitate the study of emotions. However they all share a prohibition on being published. i.e. one of their conditions of use is that they are not placed in a directory open to web access. The images in the Abstract500 are all Creative Commons licenced and this allows that image set to be used as a tool to gather design feedback via a web service. If an emotive image set to use alongside the Abstract500 in a similar way is needed, then none of these established emotion stimuli image sets will be open to this use.

Therefore, a new image set will need to be assembled specifically for the purpose.

8.3 Choosing an Emotion Model

There was concern that the model of emotion (Geneva emotion wheel) used as a source for the terms used in Chapter 7 to evaluate the summarisation method may not offer the resolution we were hoping to achieve in our visual pallet of emotions. For this reason a closer examination of emotion models was undertaken.

8.3.1 Models of Emotion

There is no agreement in the psychological literature on one model of the structure of emotion, indeed there are several models of emotion. Power (2006) summarises the existing models as belonging to 3 categories:

1) Positive or Negative: - There are positive or negative classes of emotions; i.e. emphasising the so called "valence" dimension; e.g. Watson & Clark (1992). Studies here have focussed on the conscious reporting of emotions (or affect) experienced by subjects.
2) Basic Emotions: - Theories in this category assert that there is a small set of basic emotions and other emotions are derived from these; e.g. Plutchik (1997) or Ekman (1999). There is little agreement on a specific set but there is agreement on 5: sadness, happiness, disgust, anxiety, anger. Many studies supporting this focussed on physiological measures and facial expressions, rather than conscious affect reporting. The model used to source emotion terms in Chapter 7 the Geneva emotion wheel (Scherer, 2005) sits in this category.
3) Differential Emotions: - Like the category above, this relies on a basic set of emotions. However, they are each separate with their own basis in the brain and in evolution; e.g. Izard (1971). However, there is some doubt about whether some of these emotions may not in fact be cognitive states and that some may be derived emotions.

Thus with three categories of emotion model, several different specific models and no agreement on which is best in the literature, there were a number of potential candidates.

8.3.2 Criteria for Choice of Emotion Model

The criteria considered when choosing the emotion model to use for our purpose are set out in Table 8.3.

EMCC No	Emotion Model Choice Criteria (EMCC)
1	The model must offer categories.
2	It must offer good resolution in terms of number of categories.
3	The categories should allow resolution of emotion intensity or degree.
4	The categories should be readily useable by potential participants who may be required to categorise images using the model.

Table 8.3 - Criteria for choosing an emotion model with which to categorise images.

The requirement for categories (EMCC1) rules out those models based on dimensions ("Positive or Negative" in 8.3.1). However, one model in particular has the characteristics suited to our purpose. The next section introduces the model and describes why it was deemed suitable contrasting it with the Geneva wheel model used in Chapter 7 as the source for emotion terms used in evaluating the Abstract500.

8.3.3 A Multidimensional Model of Emotion

Plutchik & Conte (1997) developed a multidimensional model of emotions that, like the Geneva emotion wheel, uses a 2D wheel or circumplex of basic and derived emotions but has a third dimension of intensity.

Evidence for the circumplex was presented by Russell (1980) by using circular ordering with polar opposites directly opposite in a circular scale (Ross, 1938). Russell (1980) used a circular ordering task along with non-metric MDS to produce a 2D layout of the initial 8 emotion categories that he investigated. This was followed by a *"category-sort task"* where 28 words were sorted into the 8 categories. What Plutchik & Conte (1997) added to the circumplex model is the subsidiary derived emotions based on adjacent basic emotions and an intensity dimension. The intensity dimension was added to account for the language of emotion i.e. various terms describing emotion. (Figure Figure 8.1).

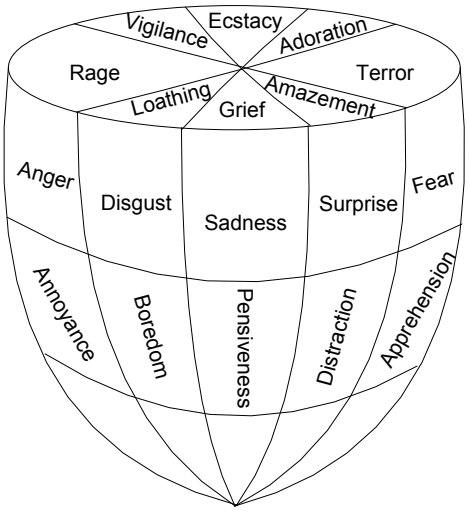


Figure 8.1 - A multidimensional model of emotions with the vertical dimension of intensity and emotion families arranged by similarity. (Adapted from Plutchik (2003)).

The emotion wheel models are formulated on similarity of their component emotions elucidated with the aid of MDS. The intensity dimension added by Plutchik & Conte

(1997) has produced an emotion model consisting of emotion terms that could be used to capture emotions and their intensity using just terms rather than terms each with a scale (as recommended with the Geneva emotion wheel (Scherer, 2005)).

The Plutchik multidimensional model would be suitable for a tagging task to categorise image stimuli by emotion terms. It offered higher resolution in the number of terms (32 discrete terms as compared to 20 in the Geneva wheel model). This would be useful in controlling the emotion content of any set derived using it. The model when opened out as in Plutchik (2003) into two dimensions, provides a circular layout in which the emotion families are presented as spokes and the intensity dimension is represented by proximity to or distance from the centre. This arrangement is easy to understand and would help in categorisation. Figure 8.2 shows this view with the addition of numbers, 1-56, and symbols, + and -, used later for emotion tagging).

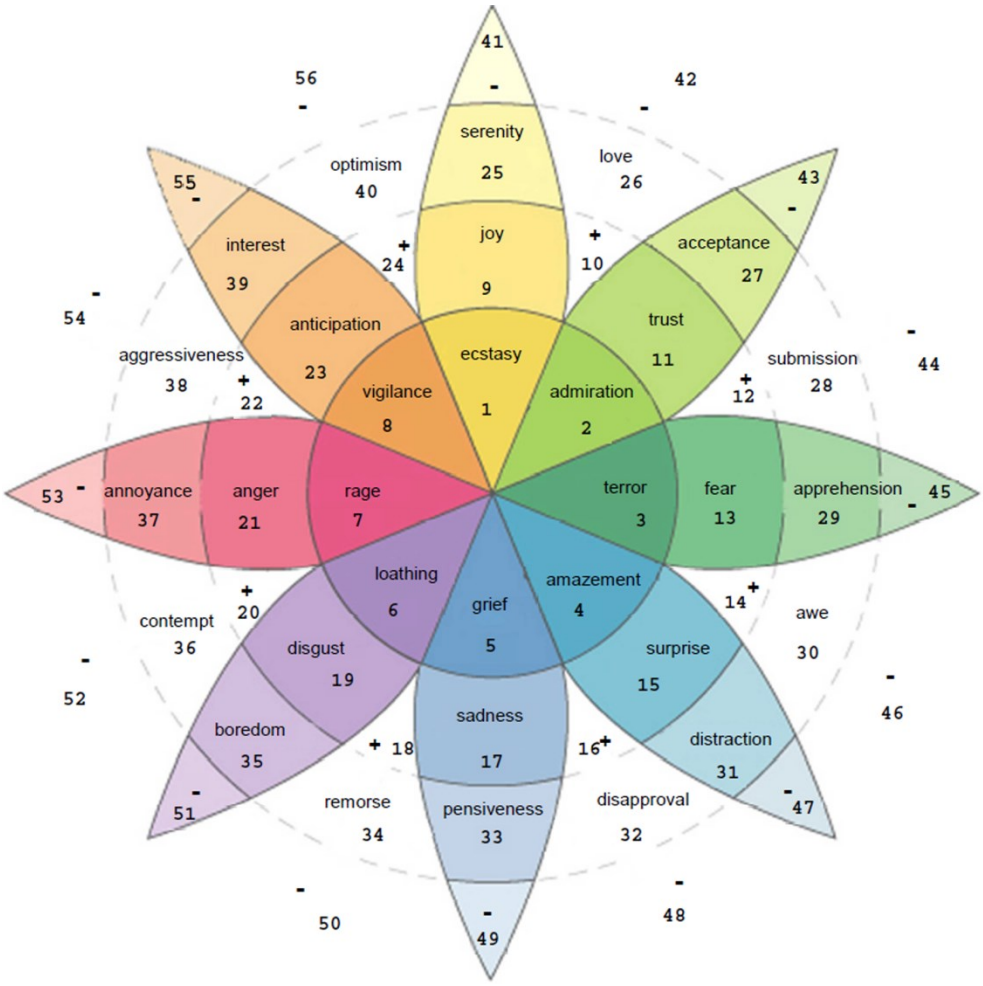


Figure 8.2- Plutchik model numbered for the emotion tagging task(adapted from Plutchik (2003)), showing the 56 tag locations; e.g the term, love, could be tagged in one of three locations indicating "love-" ,less intense love, "love", medium intensity, or "love+" for more intense love.

The Plutchik model has been used by others to inform computer interface development; e.g. Kajiyama & Shin'ichi (2014) and Cambria et al. (2012).

This model meets all four choice criteria in Table 8.3 and is therefore chosen for our purpose.

8.4 Assembling a Set of Candidate Images

8.4.1 Limiting the Scope of the New Emotive Image Set

The emotion model includes the full range of emotions. However there are emotions such as *fear* which are unlikely to be relevant to a design conversation. This fact might allow the scope of a new image set to be limited and thus reduce the size of the task and the resources required to meet it.

To allow the new image set to focus on emotions for design communication, a subset of the terms on the Plutchik model was sought. Staff and students at a design institution were surveyed. (See Appendix D p.216). As a result, 19 suitable terms were selected from the model. These included for example, *joy* and *aggressiveness* but excluded, for example, *ecstasy* and *loathing*.

Thus, a subset of 19 terms from the emotion model suitable for design feedback was defined (the *design feedback emotion subset*). These terms would be used as the basis of search terms for gathering the images. (See Appendix D p.218).

8.4.2 Gathering the Images

Images were gathered through several systematic screen scrapes. A database of search terms was constructed to support the automation of the scrapes. These search terms were based on the *design feedback emotion subset* and synonyms. (See Appendix D p.218). The scrapes were carried out so as to gather images in quantities balanced across the *emotion subset*. 5090 Creative Commons licenced images were identified and downloaded.

During an initial view (by viewing arrays of thumbnail images) 1770 were rejected on grounds of obvious repetition (often 10 or 20 images on one theme had been scraped

from a site). Using similar methods and an image database manager based on that used for the Abstract500 a further 1138 were marked as unsuitable. The criteria for this included a) distracting attribution labels b) inappropriate images which had passed the Google and Flickr "safe search" switches and c) random rejection of images associated with specific search terms for the purposes of even coverage over the 19 *design feedback emotion subset* terms.

Consideration had to be taken here of the size of the final emotion set compared to the Abstract500. Too few images in the emotion image set would risk users pecieving it as sparse compared to the Abstract 500. A minimum target of at least 200 emotion images was considered: still significanly less than 500 but still numbering in the hundreds (and an average of 10.5 per *design feedback emotion subset* term). The descision was made to aim for 200 at least in a filtered emotion image set. This is set out in Table 8.4 as a secondary image set requirements.

SISR No	Secondary Image Set Requirement (SISR)
4.1	A minimum population target of 200 images was set for the size of the emotion image set for design feedback.

Table 8.4 - Secondary image set requirement: a minimum population target..

On average 105 images associated with each of the 19 search terms were selected at random (to total 2000) from those 2182 that remained. This would allow the success rate at finding good images for any given term to be as low as 15%. This would allow at least a balanced set of, on average, 300 images to cover the *design feedback emotion subset*, well above the minimum target of 200 set in Table 8.4.

In this way the *Emotive2000* image set, a database of 2000 Creative Commons images (with accompanying attribution and search term data) suitable to be shown to image categoriser participants was assembled. The 2000 images were balanced across the 19 terms of the *design feedback emotion subset* (on average 105 per term).

8.5 Obtaining Category Data on the Emotive2000

Although the screen-scraped images were each already associated with one of the terms, the accuracy of the association by tagging with (or by co-proximity on a web page to text containing the term) was not reliable. The emotive content of the images needed to be explicitly read to allow the images to be categorised. With 2000 images to categorise it was decided to use crowdsourced participants to categorise the images. This would

allow a high volume of judgements to be collected thus increasing the overall reliability of the tagging on each image.

This section first (in 8.5.1 to 8.5.3) describes the creation of an application to allow classification of images through tagging with emotion terms and then details the steps taken to address the issue of data quality control when using the application to obtain tags through a crowdsourcing service. In 8.5.4 and 8.5.5 the formulation of the stimuli packets and setting of participant pay are described. Subsection 8.5.6 describes how the categorising application was administered on CrowdFlower limiting participation to native English speakers. Subsections 8.5.7 to 8.5.10 describe how the quality of the crowdsourced observations was assessed, how a particular quality control threshold was set, and how the effectiveness of the categorisation was evaluated in early batches of images. 8.5.11 sets out statistics describing the completed data collection exercise including the quality control rejection rate. 8.5.12 and 8.5.13 describe the construction of an emotion profile for each classified image and the assembling of the fully categorised set of 2000 images into a SOM browser based on these emotion profiles to produce an overview of the *Emotive2000* image set following categorisation. This section is summarised in 8.5.14.

8.5.1 Emotion Categoriser for Images (ECI) Application Interface

A web application was created to manage an unsupervised categorisation task (Ashby et al., 1999) allowing users to tag images with emotion terms by dragging-and-dropping them onto a version of the Plutchik wheel emotion model. The application is illustrated in Appendix D p.220. It can collect data of slightly higher resolution than simply the 32 terms on the Plutchik model. The area on wheel model was divided up into 56 tag locations (Figure 8.2). Each image classification reading consisted of the image ID and from zero to five tags, each representing one of the 56 member Plutchik emotion model tag locations. (Tag locations and terms were not permitted to be tagged twice). The application also included a database to serve randomly ordered stimuli packets (sequences of images IDs) and to record the tagging judgements.

Thus a web application, the ECI (emotion categoriser for images) facilitating drag-and-drop emotion tagging of images by remote participants was created.

8.5.2 Approach to Data Quality Control (*QC*)

As previously noted in (Chapter 4), one specific issue in employing crowdsourced participants in providing judgements is that of "cheaters"; i.e. avoiding accepting into the data, judgements from insincere participants who seek to exploit the crowdsourcing platform for unfair monetary gain.

A conventional approach to quality control (*QC*) in crowdsourcing is *the "gold set" approach* in which the stimuli for which judgements are sought are interspersed stimuli for which the correct judgements (or answers) are already known i.e. termed the "gold set" (Kazai, 2011) against which the performance of workers is assessed allowing their other judgements to be accepted as reliable or rejected as unreliable.

A less conventional approach to quality control was taken in Chapter 4 due to the inability to establish "gold set" data. Instead an approach based on a) using time on task as a criteria on which to identify possibly shoddy workmanship b) manual checking of suspect work and c) offering a bonus for more thorough work, was used.

However, this time the circumstances were different: a) It should be possible to establish a "gold set" of images with clear emotion tagging solutions against which to assess a worker's judgements, b) the required number of judgements and thus participants, would be greater than for the Abstract500 meaning the feasible degree of manual scrutiny of borderline cases would be proportionally less and c) it would not be possible to offer worker bonuses. (The project no longer had access to the AMT service, and so an alternative service had to be used. CrowdFlower was selected as it did provide indirect access to the AMT workforce. However, CrowdFlower did not allow a bonus to be offered for extra care)

For these reasons, therefore, for the emotive image set, the more conventional "gold set" approach (Kazai, 2011) would be used involving a "gold set" of images for which the correct tags would be known.

8.5.3 Establishing the Gold Set for Quality Control

A "gold" data set in the form of five images with reliable, demonstrated, emotion tag profiles (the *Gold Set*) was established by

1) Surveying 20 locally sourced participants asking them to categorise images using tag location IDs from the numbered Plutchik model (Figure 8.2) as categories or "tags". (See "Gold Set image survey" in Appendix D p.218).
2) Collating the tags from the survey results (See Appendix D p.219).
3) Using those results to produce the *Gold Set data* for quality control in the form of acceptable tagging patterns for each of the five *Gold Set images*. (See Appendix D p.219).

Thus the *Gold Set*, five images with relatively narrow ranges of associated emotion terms for assessing the quality of categorisation by crowdsourced subjects, was produced.

8.5.4 The Stimuli Packets for the ECI

Each sequence of 32 stimuli (a stimuli packet) to be tagged by participants consisted of two training stimuli followed by 25 actual stimuli interspersed with five Gold Set images. For details of how this make-up was arrived at and how the stimuli packets were generated see Appendix D p. 220. The stimuli packets were constructed such that images could be fully classified (with 20 "readings" per image) in batches of 100 to allow early and periodic assessment of the effectiveness of the process with defined random batches of images from the *Emotive2000*.

Participants were discouraged from doing more than one stimuli packet by a) the task instructions and b) using cookies in the application.

8.5.5 Participant Pay

For details of the consideration given to what to pay participants see Appendix D p.221. After due consideration pay was set at $1 per HIT with one stimuli packet per HIT (Human Intelligence Task (Kazai, 2011)).

8.5.6 Running the ECI Application on CrowdFlower

Aside from specifying the pay and task details, the CrowdFlower interface permitted providers to a) select worker channels, e.g. AMT and Entropia Partners (2015) and b) choose to offer the HITs in a restricted list of countries.

The ECI HITs were restricted to countries where English is the native language i.e. Australia, Canada, Falkland Islands, United Kingdom, Ireland, Isle of Man, New Zealand, United States. This was so as to avoid misunderstandings of the emotion tags. The HIT instruction also stated the task should not be attempted by non-English speakers. This will inevitably introduce some cultural bias into the image set. The additional complexity and expense of designing image sets without cultural bias is hereby set outside the scope of this thesis. However this issue, with particular reference to validity, is discussed in the final chapter. A number of worker channels were selected but most of the HITs were done by AMT and Second Life workers. (See Appendix D p.221, for details of running the HITs on the CrowdFlower service).

Thus, the HITs on CrowdFlower were restricted to native English speakers.

8.5.7 Assessing the Quality of the Crowdsourced Tags

A properly completed HIT produced one *set of observations* by one participant. A *set of observations* consisted of zero to five tags per image given to the image stimuli in a stimuli packet. **The reliability of a set of observations was assessed by *comparing*, the tagging of the five Gold Set images within that set of observations**, ***with*** **the Gold Set data**. That set of observations was given a quality control score (*QC score*). The QC score for each set of observations would later be compared to a Quality Control threshold (*QC threshold*) when deciding whether to accept or reject that set of observations. Equation (8.1) shows the calculation of a QC score for a set of observations, in which the QC score equals the sum of the component scores for each of the five Gold Set images, $i_{1..5,}$ where n is the number of tags given to Gold Set image i, x is the number of hits (i.e. acceptable tags), and y is the number of misses (i.e. unacceptable tags). (See Appendix D p.222 for details of how this was done).

$$QCscore = \sum_{i=1}^{5} \frac{1i_x + 0i_y}{i_n} \qquad (8.1)$$

8.5.8 Requirements for the Quality Control Threshold

Table 8.5 sets out the criteria used when setting the quality control threshold.

QCTC No	QC Threshold Criteria (QCTC) The threshold should allow rejection of observations from
1	participants who submitted low quality unreliable data without the wastage of observations from participants who provided good quality reliable data.
2	participants who did not sincerely attempt the task and provided nonsense tags.
3	participants who over tagged, by trying too hard and tagging with the maximum possible tags thus diluting the good tags with unreliable ones.

Table 8.5 - Criteria for setting the QC threshold.

8.5.9 Setting the Quality Control Threshold

The quality control threshold had two objectives:

1) To identify stimuli packets associated with low QC scoring sets of observations so that further sets of observations could be sought for those stimuli packets.
2) Allowing the eventual collating of all sets of observations associated with QC scores over the threshold to produce a quality controlled set of observations to use for final results.

The ECI database consists of linked tables and allows sets of observations to be sampled based on their QC score by running queries. By sampling sets of observations at various levels of QC scores, the score of 3.1 out of 5 was set as the threshold below which the stimuli packets associated with such sets of observations would be "recycled", i.e. made available again to the ECI app so that another set of observations would be sought to satisfy that stimuli packet. The rejected sets of observations were not discarded. They remained in the database and would be available for analysis. (See Appendix D p.223 for details of how this was done).

In summary: a QC threshold was established such that, using this threshold, a) further observations could be sought for some stimuli packets, and b) the reliable sets of observations could be extracted from the ECI database as results.

8.5.10 Evaluating Effectiveness of Tagging in Early Batches

When tagging of the early batches of images was completed, steps were taken to evaluate the success and validity of the tagging. These steps consisted of a) developing charts to visualise the emotion profile of each image (Figure 8.3) b) assembling the tagged batches into SOM browsers to check that tagging produced sensible stacks/clusters and c) interrogating the structure of the developing image set using an

interactive dendrogram application. Details of how this was done are in Appendix D p.224.

These evaluations lead to the conclusion that 500 images did not contain enough images with clear peaks (in their emotion profiles) for some terms (in particular for "Aggressiveness" and "Disapproval"). Thus it was decided to proceed with tagging the whole 2000 images in the expectation the remaining 1500 untagged images included enough of the "missing" categories.

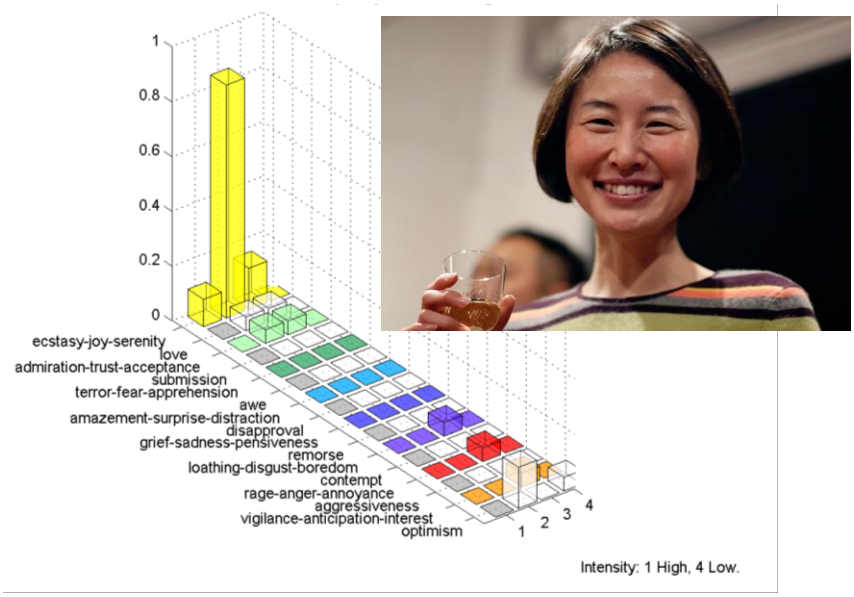


Figure 8.3- Image ID103 (inset) and its quality controlled emotion tag frequency vector viewed as a chart. The chart represents the Plutchik emotion model used for the tagging (Figure 8.2) "unzipped" down the ecstasy/joy/serenity spoke (or emotion family) and opened out. The y- axis shows the normalized tag frequency; the emotion spokes are arranged along the x-axis; the intensity radii are labelled 1 to 4 on the z-axis, 1 being most intense. Cells coloured grey are null as there are no tagging locations on the emotion map at these places.

8.5.11 The Finalised Results Data Collection Statistics

Percentage of observation sets accepted/rejected		
Description	**Quantity**	**%**
Total full sets of observations recorded	1972	100
Sets of observations which passed the QC threshold into the results.	1605	81
Sets of observations which failed QC threshold and were rejected.	367	19

Table 8.6 - Quality control rejection rate: sets of observations accepted and rejected.

Once all the stimuli packets had been completed by a set of observations which had passed quality control data collection was ceased. An evaluation of the data collection

operation was carried out by collating relevant statistics. These are set out in Table 8.6 to Table 8.9.

Opportunities			
Counts		**Statistics**	
No. of Opportunities	**Count**	**Median**	20
19	8	**Mean**	20.07
20	1841	**Max**	21
21	151	**Min**	19

Table 8.7 - Tagging opportunities count for the Emotive2000 images (used for normalising the tag frequencies).

Participants	
Total number of individual participants in final accepted data set	905
Total sets of observations in final accepted data set	1605
Accepted sets per participant	
Mean	1.8
Median	1
Mode	1
Max	17

Table 8.8 - Participant statistics for the observation sets in the final accepted data.

Cost	
Total expenditure including pay and CrowdFlower commission	£1777.00
Cost per image	£0.89

Table 8.9 - Cost of the crowdsourced data collection.

Table 8.9 shows that the cost of the data collection when considered per image (£0.89) was quite reasonable.

As stated in 8.5.4 participants were discouraged from doing the task more than once. Indeed Table 8.8 shows that the typical participant did do it once only, but one particularly tenacious participant did the task 17 times. However, the images were tagged in batches and the probability that the same image was tagged twice by the same participant is low as, by the time a participant would be able to repeat a task, it is likely that this would be in a new batch. In addition, Table 8.7 shows that the vast majority of the images were tagged 20 times with only 8 out of 2000 being tagged less (19) times. Thus it is safe to describe the Emotive2000 profiles as representing the judgments of 20 individuals.

It can be seen from Table 8.6 that the quality control formula rejected 19% of completed observation sets.

8.5.12 Building the Emotive2000 Emotion Profiles

The emotion profiles for all of the Emotive2000 were assembled from the quality controlled tag frequencies. (See Appendix D p.225 for details). The emotion profile components for each image are listed in Table 8.10.

Emotion profile component	Description
Tag frequency vector	56-member vector defining the tag location frequencies normalised by the number of tagging participants. (Figure 8.3)
Term frequency vector	32-member vector representing the tag frequency vector collapsed down to the terms. This is also normalised for the number of tagging participants. (Figure 8.4.)
Tag and term frequency vector charts	Charts of the vectors laid down on the emotion model.

Table 8.10 - The components of each image's emotion profile in the Emotive2000.

Figure 8.4 illustrates how the tag locations on the emotion model, and the tag frequency vectors, could be collapsed to form term frequency vectors. Notice in the figure that on the left "love" is represented by three tag locations (love -, love, and love +), while on the right those three tag locations are aggregated into one term, love. The terms such as acceptance which have 2 locations (acceptance and acceptance -) are likewise aggregated in the term vector. The emotion intensity z-axis is not labelled in the term vector chart as each term has only one intensity in that collapsed view.

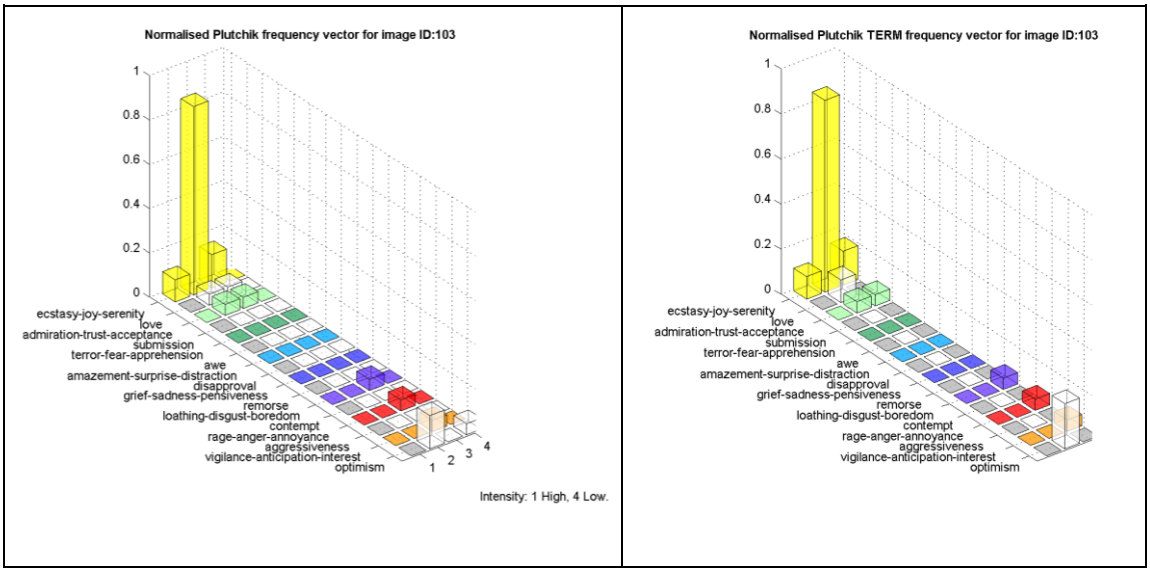


Figure 8.4 - The tag frequency vector (left) for image ID103 with the corresponding term frequency vector (right).

8.5.13 The Emotive2000 Image Set in a SOM Browser

The full 2000 images, characterised by their quality controlled emotion profiles were assembled into a SOM browser (Figure 8.5). (See Appendix D p.226 for details). Constraining the dimensions to 9x7 stacks produced a browser that is usable with a relatively large monitor. The images can be browsed and clicking a thumbnail in an open stack displays the image database record with full size image and both tag and term frequency vector charts.

Figure 8.5- Full Emotive2000 in a 9x7 stack SOM showing top level (left) and three open stacks (right). The ECI images database record (opened when a thumbnail in a stack is clicked) is also shown (bottom left). This is a composite of screenshots from a web browser.

8.5.14 Summary of Section 8.5

Section 8.5 described how reliable human derived category data, on each image in the Emotive2000 image set, was obtained using a crowdsourced tagging application. A "gold set" (Kazai, 2011) approach to quality control was adopted. A *Gold Set* of images with known emotion profiles was established and these were included within the stimuli packets to be tagged. All sets of tagging observations were then given a quality control score (*QC score*) by using a formula which compared the tagging of the Gold Set images with the Gold Set known emotion profiles. A QC score threshold was set such that unreliable sets of tagging observations could be rejected. (In fact 19% of observation sets were rejected by the formula based on their QC score.) Reliable sets of

tags were collated and emotion profiles for each image were produced, each, representing the tagging judgements of 20 individuals. An emotion profile comprises two vector formats, a 56-member tag vector and 32-member term vector, along with chart visualisations of both vectors. Lastly the Emotive2000 was assembled in a SOM browser to permit an overview of the set and the viewing of each image with its profile.

8.6 Filtering the Emotive2000 Image Set

This section addresses two of the requirements set out at the start of the chapter in Table 8.1., specifically PISR 1c) and 4; i.e. that the image set if possible have an even spread of emotions suitable for design feedback and that there must be enough to offer users a wide choice and enough that selections drawn from the set merit summarisation.

There was an imbalance in the representation by term within the image set. Figure 8.6 shows this imbalance graphically by illustrating the number of images whose highest profile peaks represented each of the 19 *design feedback emotion subset* (established in 8.4.1). There was a need to filter the image set to avoid images representing some terms being over-represented risking bias in the feedback generated using the image set.

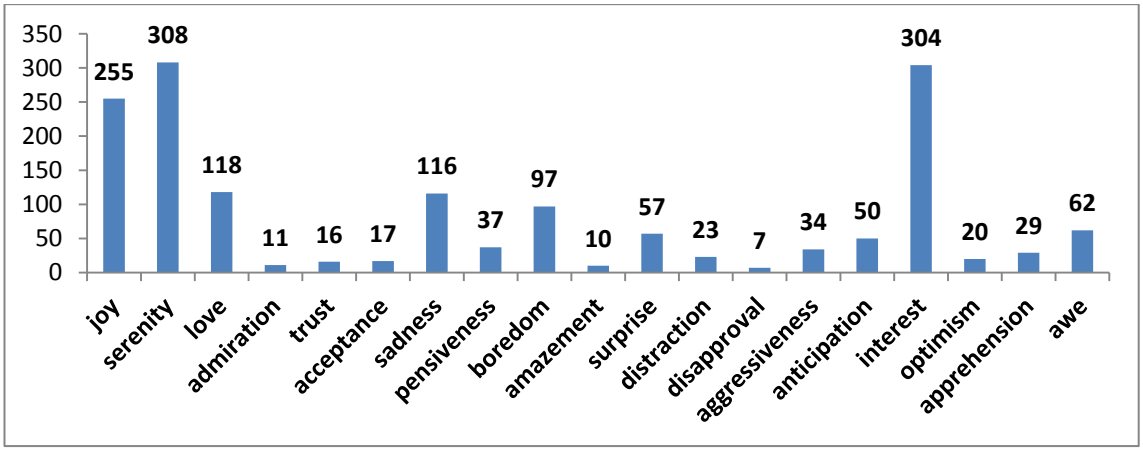


Figure 8.6 - The number of images in the Emotive2000 ranking first by search term. The chart shows the number of images having their highest normalised tag frequency peak in their emotion profile associated with the design feedback emotion subset terms.

Simply judging an image's meaning by taking its highest emotion profile peak was a crude measurement as images often had several peaks. However, measured this way, the lowest representation in the set was for the term, *disapproval*, (just seven images, see Figure 8.6). This suggested it might be possible to take seven images per term but this would only then allow a set of 7x19 = 133. The minimum population target set out in

Table 8.4 was at least 200 emotion images an average of 10.5 per term. The next lowest represented terms were, *awe*, with 10 images, and *admiration* on 11, by the measure in Figure 8.7. With a target of 200 only 1 term (*disaproval*) would be under represented, by the measure used in Figure 8.8. The decision was made to continue aiming for a population of 200 at least in a filtered emotion image set.

Thus, the image set was filtered to produce the best 200 images (at least) for the *design feedback emotion subset*. This set contained 204 images (*Emotive204*). (In fact the filtering used was more nuanced than the measure charted in Figure 8.6 and also took account of the contrast between term peaks within images' emotion profiles. See Appendix D p. 227 for details).

8.7 Assembling Emotive204 in a SOM Browser

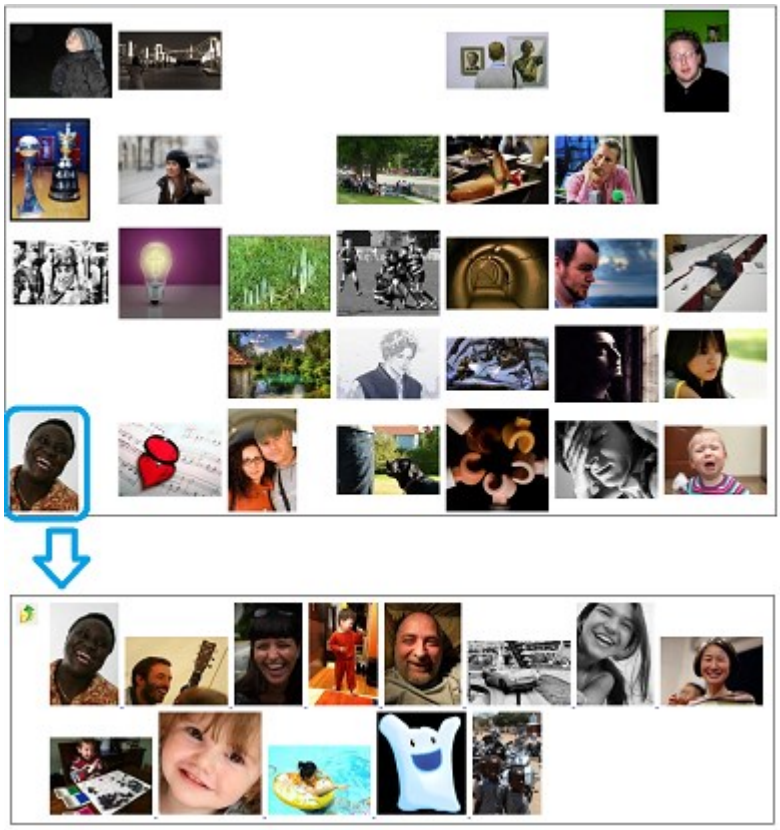


Figure 8.9 - Emotive204 in a 7x5 stack SOM showing top level (top) and an open stack (bottom)

Although the *Emotive204* filtering was based on the 32-member term vectors it was assembled it in a SOM browser (Figure 8.9) based on those 204 images' 56-member tag frequency vectors to make use of the higher resolution data represented by the added

intensity discrimination available to tagging participants for terms such as '*love'* with three tag locations and '*serenity'* with two tag locations on the emotion model as offered to tagging participants (Figure 8.2).

8.8 Conclusion to Chapter 8

This conclusion begins with an overview of the chapter and then the image set requirements which were established at the start of the chapter are revisited.

8.8.1 Overview

Having established in Chapter 7 that a new image set to enable emotion communication was required and set out the requirements for such an image set, emotion models from the literature were discussed. One (Plutchik, 2003) was chosen as suitable for our purpose based on criteria set out in Table 8.11. A major deciding factor was its use of the language of emotions and its intrinsic emotion intensity dimension embedded in a set of 32 emotion terms. A subset of these terms, *design feedback emotion subset*, was identified as being suitable for this research.

2000 Creative Commons images associated with this subset of terms were gathered. To obtain high resolution data describing the emotion content of the images paid crowdsourced participants tagged them using a drag-and-drop interface with 56 tag locations based on the emotion model. Quality control was based on a *gold data set* of images with tagging established by lab based participants. The output of the crowdsourced categorisation was *Emotive2000*, a database containing 2000 images with emotion profiles. Each image's profile consists of a) a high resolution 56-tag frequency vector b) a lower resolution 32-term frequency vector and c) visualisations of both vectors charting the frequencies laid out on the emotion model. The *Emotive2000* data set is accessible via a SOM browser combined with a database containing attribution details and search terms used to gather each image.

Emotive2000 was further filtered to produce *Emotive204*. This smaller set of images was balanced across the *design feedback emotion subset*. *Emotive204* was assembled in a SOM browser based on its emotion profiles.

8.8.2 Image Set Requirements revisited

ISR No	Image Set Requirement (ISR)	Met?
1	The set must a) be communicative of emotions b) those emotions should be suitable for design feedback; and c) if possible an even spread of emotions should be sought to reduce the risk of biasing the feedback	Yes*
2	Data must exist (or be obtainable) on each image, suitable to allow a) deployment of the image set in a SOM browser, thus permitting user interaction similar to the Abstract500 in the SOM browser; b) summarisation of selections from the set; and c) a degree of control over the emotion content in the set (to help with 1 b) and c) above).	Yes
2.1	The image data should include a form of emotion categorisation.	Yes
3	The images must be free to use.	Yes
4	There must be enough images in the set to offer users a wide choice and enough such that selections drawn from the set merit summarisation.	Yes
4.1	A minimum population target of 200 images was set for the size of the emotion image set for design feedback.	Yes

*Table 8.12 - Requirements for an emotive image set revisited. The Primary and Secondary requirements (from tables Table 8.1, Table 8.2 and Table 8.4) are combined in this table. *See caveat in text following this table.*

There is a caveat to the labelling in Table 8.12 that Requirement 1 has been met: Compromises were made in the filtering producing Emotive204. Several of the *design feedback emotion subset* terms were represented by images with low frequency peaks for that term despite it being the highest frequency peak in that image's profile. However, there was not time to refine the filtering method further prior to the CVFM evaluation studies and thus development of the filtering ceased as it stood. It remains to be seen how effective Emotive204 will be for design feedback.

Thus, in summary, this chapter has produced

1) *Emotive2000*: A database of 2000 Creative Commons images with emotion profiles consisting of two emotion vectors (one lower and one higher resolution). This data set, although actually a by-product, may well be of use in other research beyond this thesis such as in training a classifier in machine learning or as ground truth data for evaluating an emotion recognition feature set.
2) *Emotive204*: A subset of the above which is balanced over a set of *design feedback emotions*. It is deployed in a web-based SOM browser and is suitable for use in evaluating the CVFM.

Chapter 9

Evaluation of the CVFM – Study Design and Pilot

Having developed and built the abstract and emotive SOM image browsers (see Chapter 4 and Chapter 8), developed the summarization algorithm (Chapter 6) and shown that summaries of image selections taken from the abstract browser are as effective at communicating terms as the image selections themselves (Chapter 7), we are now in a position to evaluate the CVFM from a user perspective.

The purpose of this chapter is to

1. Describe the aims of the evaluation with rationale.
2. Describe the methodologies and study design used and the rationale for their selection.
3. Describe a pilot of the evaluation study and the results of that pilot.
4. Discuss the pilot study results and effectiveness of the methods used, concluding with modifications to the study design and methods for use in the main evaluation study.

Appendix B is the appendix associated with this chapter.

9.1 Aims of the Evaluation Study

The overall aim is to establish the viability of the image feedback method from a user perspective. Essentially we are interested in whether or not users will be engaged by and value this way of communicating. Table 9.1 specifies this issue with a reference number for use later in forming research questions.

No	User Issue	Priority
U(i)	Users are engaged by and value the method	High

Table 9.1 - User issue for evaluation

In addition to this overarching aspect, some specific issues of the shortcomings of conventional feedback methods, cognitive psychology and emotion in design were discussed in Chapter 2 as being part of the motivation in developing the CVFM. These are summarised below in Table 9.2 along with subsection references. A priority was assigned to each issue to help in scoping the study.

No	Motivation Issue	Priority	Sub-section reference
M (i)	Visual-verbal dimension	High	2.6.1
M (ii)	Intuition	Med	2.6.2
M (iii)	Emotion expressiveness for design communication	High	2.7.1
M (iv)	Selective non-response	High	2.1.1
M (v)	Social desirability response bias	High	2.1.1
M (vi)	Overly negative responses	Med	2.1.2
M (vii)	Contributors arguing for their opinion	Low	2.1.2
M (viii)	Expression of emotions discouraged	Low	2.1.2

Table 9.2 - Motivation issues for evaluation

The first four issues listed in Table 9.2, M (i) to M (iv), overlap; e.g. we expect that the cognitive style (on the visual-verbal dimension) within feedback givers may affect the eventual response profile (including non-response) of the method.

9.1.1 Research Questions

Candidate research questions, set out in Table 9.3, below, were framed to address the issues identified above in Table 9.2.

One issue not addressed by a research question is M (vii) i.e. contributors arguing for their opinion. This will be addressed indirectly in that all feedback participants will be given equal weight in the study.

ERQ9 (evaluation research question 9) is open ended to take account of the possibility that designers may hold unforeseen views about the visual feedback formats.

ERQ No	Evaluation Research Question	Issue ref.
Feedback givers (the crowd)		
1	Do feedback givers prefer using images or text when describing their emotions?	U(i), M (iii), M (iv)
2	Do feedback givers find the image formats more engaging than text?	U(i), M (i), M (iv)
3	Do feedback givers feel able to express their answer using the image formats?	U(i), M (viii)
4	Do feedback givers feel more or less inhibited in expressing their emotions using images compared to text?	M (v)
Designers (those consuming the feedback)		
5	Do designers value the image feedback formats?	U(i), M (i)
6	Do designers prefer receiving feedback about emotions using the image formats or text as the medium?	U(i), M (i), M (viii)
7	Are designers inspired by the visual feedback to make changes to their designs?	M (iii), M (ii)
8	What do designers think of the image formats as a method of feedback about emotions experienced by viewers of their designs?	M (iii), M (viii)
9	What do designers think of this method of communication?	U(i)
10	Would designers use a service providing the visual feedback formats?	U(i)

Table 9.3 - Evaluation research questions with references to the issues which motivate them.

9.1.2 The Two Sides of the Study

The research questions reveal that there are two aspects (or sides) to the study: the feedback side and the designer side. These are defined below. Following this definition the planning, methods, and results for the two sides of the study can be addressed separately when required.

On the feedback side, participants representing the crowd (feedback participants) respond to designs as stimuli thus generating feedback of different types for the designers to view later. During this process the feedback participants will also provide data about their perceptions of using the various feedback formats.

On the designer side, the designer participants view the feedback and provide data about their perceptions, as the ultimate consumers, of the feedback.

9.1.3 Scope of the Study Related to the CVFM

The stages of the method being evaluated

How the two sides of the evaluation relate to the CVFM as set out in Figure 1.1 is described below.

1) The feedback side corresponds to stages 2 and 3 in Figure 1.1; the crowd views a design along with a question and responds by selecting images from a browser.

2) The designer side corresponds to stage 6 of Figure 1.1.

Stages 1, 4 and 5 of Figure 1.1 are not being evaluated here. Stage 5, the visual summarisation, was evaluated from a semantic effectiveness perspective in Chapter 7. Stage 1, the submission of a design by the designer, while still necessary to enable the evaluation, is done "off-line" in an administrative way via email communication with the designer participants. Stage 4, the collation of the feedback, is an administrative step which, while necessary to allow the other steps is not being evaluated.

Co-design cycles

One aspect of the CVFM is its potential for cycles of co-design to allow designers to develop a prototype through iterations to a finished design. This evaluation, however, will only seek to evaluate the method for a single cycle i.e. designers showing designs, a crowd giving feedback and designers viewing the feedback (c.f.1.3.2).

9.2 Overview of Study Design

9.2.1 Study Format

In this subsection the considerations involved in devising the study format are described.

To allow comparison with a "ground truth" condition, text feedback was to be gathered in addition to the image feedback. Furthermore, with this study there is an additional ethical consideration in that data from one group of participants will be shown to another i.e. feedback participant data in the form of image selections and text comments will be shown (anonymously) to designer participants. The designer participants have a

personal stake in the feedback. With the inclusion of text feedback there is a risk of exposing designer participants to potentially hurtful text feedback. The image feedback, being restricted to images in the image sets is a known quantity. This issue is a factor taken into consideration below.

Three options were considered

A. Live end-to-end web application allowing design images to be uploaded, feedback (including text and image feedback) collated and feedback viewed by designers, all unsupervised and with other data gathered during the process.
B. Offline collection of designs from designer participants; feedback participants recruited individually and providing feedback in individual sessions; feedback collated offline; and finally feedback shown to designers individually.
C. As for B but feedback participants recruited as a class and providing feedback in a single session.

Op-tion	**Pros**	**Cons**
A	The feedback and designer participants would experience a system close to the final envisaged system.	Not enough time available to develop the web application to integrate both a reliable user interface and the clustering. Text feedback might require moderation (Ethically, exposing designer participants to un-moderated text feedback will be problematic). Moderation would probably not be possible in live feedback. Reliable en-mass computing facilities would be required for a coordinated session. Interviews with all designers while they received the feedback not possible.
B	Only the interface to gather and store feedback from feedback participants need be developed. Existing offline clustering code can be used. Existing summary rendering application can be used. Previously used participant recruitment methods can be used.	Recruiting feedback participants for individual sessions would require a period of time (perhaps 1 week or more). Thus increasing the time between designers providing their design and receiving feedback.
C	Only the interface to gather and store feedback from feedback participants need be developed. Existing offline clustering code can be used. Existing summary rendering application can be used. A shorter time (relative to B) between a designer providing a design and receiving the feedback.	An alternative recruitment and administration policy will be required for the feedback participants. (However, a class of undergraduates was available for recruitment.)

Table 9.4 - Evaluation study format options with pros and cons.

Format A was ruled out due to the time cost in developing an integrated study and feedback application.

A major aspect of the study was to seek the views of the designer subjects. It was judged that having too long a gap between the designers submitting their designs and receiving the feedback might affect the results**.** It was, therefore, decided to opt for Format C which minimised the time between design submission and feedback.

A mixed method approach

It was decided to adopt a qualitative approach with the designer participants and use semi structured interviews to pursue their views as this would allow the opportunity to probe any unforeseen topics raised by those designers. However, as a relatively large number of feedback participants were likely to be required (too many to interview) it was decided to gather data from feedback participants, sufficient to allow a quantitative analysis of their views, at the time they give the feedback.

9.2.2 Participants

The decision was made to recruit the participants from a contextual studies class of 3rd year undergraduates. The class available to the project contained a small group of interior designers, and the remainder group was approximately 50 in size. The gender imbalance inherent in seeking participants at the textile and design campus was present in the feedback group (In the end, data was successfully collected from 32 feedback participants including just one male). However the designer group (of 12) had 3 males. The participants all received course credit for taking part. The designer participants received their choice of 100g chocolate bar as an additional thank you on completion of their interview in recognition of their additional commitment in providing a design image and booking their interview appointment.

9.2.3 Feedback Task

Feedback participants would be shown design images, asked a question and then asked to respond using the different response formats. They would be asked for judgements about each response format. Thus both feedback to fuel the designer side evaluation and data about the feedback participants view of the response formats would be gathered.

VAS items would be used for feedback participant judgments as they produce interval data (Reips & Funke, 2008) and allow parametric tests (McCrum-Gardner, 2008).

The stimulus question

As the CVFM was hoped to encourage expression of emotion, the stimulus question was worded with that in mind. The wording chosen was:

> "*How did the design make you feel?*"

Thus, feedback participants would be asked to respond to images of designs and the question, "*How did the design make you feel?*".

9.2.4 Designer Interviews

A semi-structured style (Kvale & Brinkmann 2009) was chosen as likely to provide flexibility in exploring themes that might emerge especially as there was unlikely to be time for follow-up interviews. The designers would view their feedback in the different formats and their reactions and opinions would be sought.

9.3 Feedback Side Variables

Independent variable

The independent variable was to be the method of response (*response format*) for which there would be three conditions:

1. Enter text in a text field
2. Choose three images from the abstract images browser
3. Choose three images from the emotive image browser

Dependent variables

Three things would be measured:

- Utility of response method: Did the subject feel enabled to express themselves fully?
- Degree of awareness of social desirability response bias: Did the subject feel free to express themselves?
- Engagement: Did the subject enjoy that method?

Sources of variability

The sources of variability in any collected data are discussed in Table 9.5 along with how these will be addressed.

Source of variability in feedback task	**Addressed**
Participants' unfamiliarity with the scales causing them to recalibrate their views about the extremes as they encounter the different conditions.	Training phase during which participants encounter all three conditions. Training phase readings excluded from the analysis.
Order of presentation of design stimuli.	Randomise this.
Order of presentation of conditions (text, abstract image set, emotive image set).	Randomise this (but balance this over the trials to minimise any cumulative difference).
Differences between subjects (participants).	Repeated measures design. All subjects see all the conditions.
The different design stimuli.	In the pilot show all stimuli to all participants. (there are only 5 designs for pilot). In the Main study (12 designs) randomise the designs (balanced across the trials).

Table 9.5 - Sources of variability in the evaluation study and mitigation.

A training phase would be included to familiarise the participants with the items such that they could calibrate their responses across the VAS items and all the conditions in their own minds prior to the experiment phase.

A power analysis using *G*Power 3.1* software (Faul et al., 2007) indicated that the expected 50 participants would be enough to expose a result from a large effect (r>=0.5) but not necessarily from a medium effect (r>=0.3) when running the anticipated ANOVA statistical tests.

9.3.1 Feedback Side VAS Item Wordings

Utility of response method (Utility)

This was an idea that would be relatively straight forward for participants to gauge and self-report. A simple wording was used. (See Table 9.6)

Social desirability bias (Freedom)

There are recognised to be two dimensions to social desirability response bias, namely "self-deception" and "other-deception" and there a number of scales used to measure these (Nederhof, 1985). Measuring the bias has been addressed by a) measuring these tendencies as traits in individuals and in demographics and b) in relation to specific

issues or domains (Randal & Fernandes, 1991). Our purpose differed from this in that we were interested in participants' perceptions of their freedom to express themselves in a particular medium compared to others. Indeed, it was decided that it would be outside the scope of this thesis to develop a scale to measure any social desirability bias present in responses consisting of images, which is what would be required to allow it to be compared to the bias in text responses (for which techniques already exist). For these reasons it was decided that, in this pilot, simply asking participants to consider the issue directly and self-report using a straight question would be attempted. (See Table 9.6)

Engagement

Engagement is often measured with several questions in a scale. Webster & Ho (1997) used a scale of 15 items when researching audience engagement in multimedia presentations. However, half of the items in that questionnaire addressed influences on engagement. The remaining items addressed three subsidiary aspects of engagement: "*attention focus*", "*curiosity*", and "*intrinsic interest*". In this repeated measures experiment the participants would be asked to provide judgements about the three types of response format in addition to using those response formats to react to a number of designs. Multiple measurement items would make the amount of work required of each feedback participant to be too great. It was decided to compromise and only measure one aspect of engagement, "intrinsic interest". The two items addressing this in the Webster and Ho questionnaire were "The presentation medium is fun" and "The presentation medium is engaging". The former was judged to be more suitable. It was converted to a style suited to a VAS item with opposing anchors. (See Table 9.6).

The chosen item wordings

VAS Item Wording	Anchor1	Anchor2
Measure: Utility of response method (Utility)		
How well were you able to express yourself?	Completely	Not at all
Measure: Degree of awareness of social desirability response bias (Freedom)		
In relation to freedom of expression i.e. freedom to say whatever you wanted without caring what anyone, including the designer, might think about the answer you gave: How free did you feel in giving your answer?	Totally free	Totally inhibited
Measure: Engagement (Interest)		
How interesting was this way of giving your answer?	Very much fun	Very much boring

Table 9.6 - Pilot VAS item wordings.

Table 9.6 shows the wordings of the VAS items.

9.4 Pilot Study: Initial Considerations

The reasons for piloting and overarching issues concerned with the pilot are set out here:

a) To check empirically that the chosen measurement VAS items work i.e. they make sense to feedback participants.

b) To check empirically that the feedback participant VAS items produce data that can be successfully analysed.

c) To trial the semi-structured designer participant interview format.

9.5 Pilot Participant Recruitment and Study Conditions

Participants for the pilot were recruited from same undergraduate year group as intended main study participant class, but from outside that class so as not to compromise the naïve status of main study participant class.

9.5.1 Designer Participants

Designer participants were approached and asked if they would contribute an image of one of their designs to aid our research into design feedback. Five designers who were approached provided their email address. Of those five, when subsequently contacted by email, three provided design images (two donating two images each and one donating a single image) and gave informed consent via email. All three were female. At that stage as a) the focus of the pilot was on the feedback participant task and b) it was not thought there would be time to collate the feedback and conduct pilot designer interviews, the designers were given the expectation that they would not see the feedback. (Later, time was found to collate the feedback for, and interview, one designer participant). The designers were not offered any inducement or reward for donating design images; however one designer was paid £30 in Amazon vouchers later to attend an interview.

9.5.2 Feedback Participants

Feedback participants were approached while working in a large open-plan garment production workshop at the TEX campus. They were offered a reward of their choice from a selection of 100g chocolate bars. There were 10 feedback participants, all female. This was a gender balance similar to that in the main study group.

The workshops offered a quiet spacious location and there were free work tables allowing the participant to step a few yards away from their work area to a work table nearby and do the task.

Participants were briefed about the study by the administrator (the Author) following a script. This included informing them that the designers whose designs they would view would see the comments (both visual and textual) but would not know who gave them. At the end participants were debriefed following a script. The purpose was to inform them that it was unlikely that all the designers would get to see the comments, that this had been a pilot mainly to try out the feedback participant task, and that it had been necessary for them to believe that the designers would definitely see the comments to allow them to properly address the "Freedom" item questions. The briefing script, debrief script, and an example task/questionnaire can be seen in Appendix B.

9.6 Feedback Side Task

9.6.1 Interface and Recording Method

A simple web interface was constructed to allow access to stimuli and to the three response formats. This could be used on a laptop with a mouse. Screens from the interface can be seen in Appendix B (p.188). The web application stored the responses in hidden page elements. After a session had finished the responses could be copied and saved in a text file.

A participant task/questionnaire sheet stepped the participant through the task, prompting them to use the different parts of the interface and record their progress and answers to questionnaire items as they went. (See Appendix B p.184).

The response formats were labelled with randomly selected letters so as to avoid introducing any preconceptions (of precedence or concepts) into the minds of the

participants. Abstract images, emotive images, and text were labelled L, P, and Q respectively.

The VAS items were implemented on paper. The results would be processed by measuring the distance to the nearest 0.5 mm from the left hand anchor. The resulting number would be recorded on the page and could then be entered into a spread sheet for processing. This avoided consuming time to develop a software application with a database purely for the pilot.

9.6.2 Training Phase

The training phase was to allow participants to experience all three answer formats first. This was so that they had the full context in mind before they were tasked to use the first of the VAS items and would be able to interpret the VAS item anchors in the light of that full context.

The administrator explained the two work flows (training phase and experiment phase) to each participant at the start of the task. The administrator then sat far enough away so as not to inhibit the participant but close enough for the participant to feel able to easily ask for guidance. (This was typically 10 feet away and at a different table). The training phase work flow for feedback participants is shown in Figure 9.1.

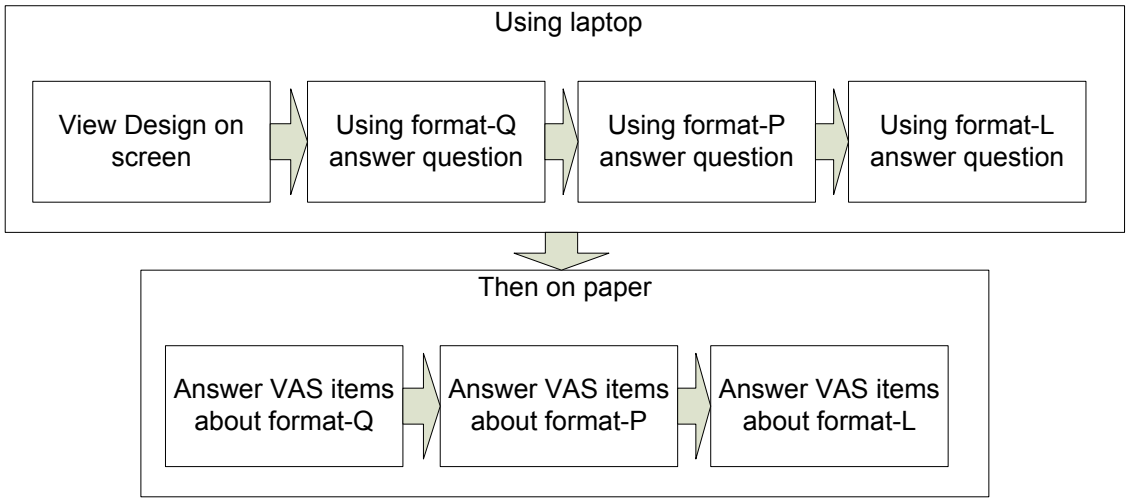


Figure 9.1- Pilot evaluation training phase workflowfor the feedback participant task. The presentation order of formats was randomised (See page 188)

9.6.3 Experiment Phase

The training phase was immediately followed by the experiment phase (Figure 9.2).

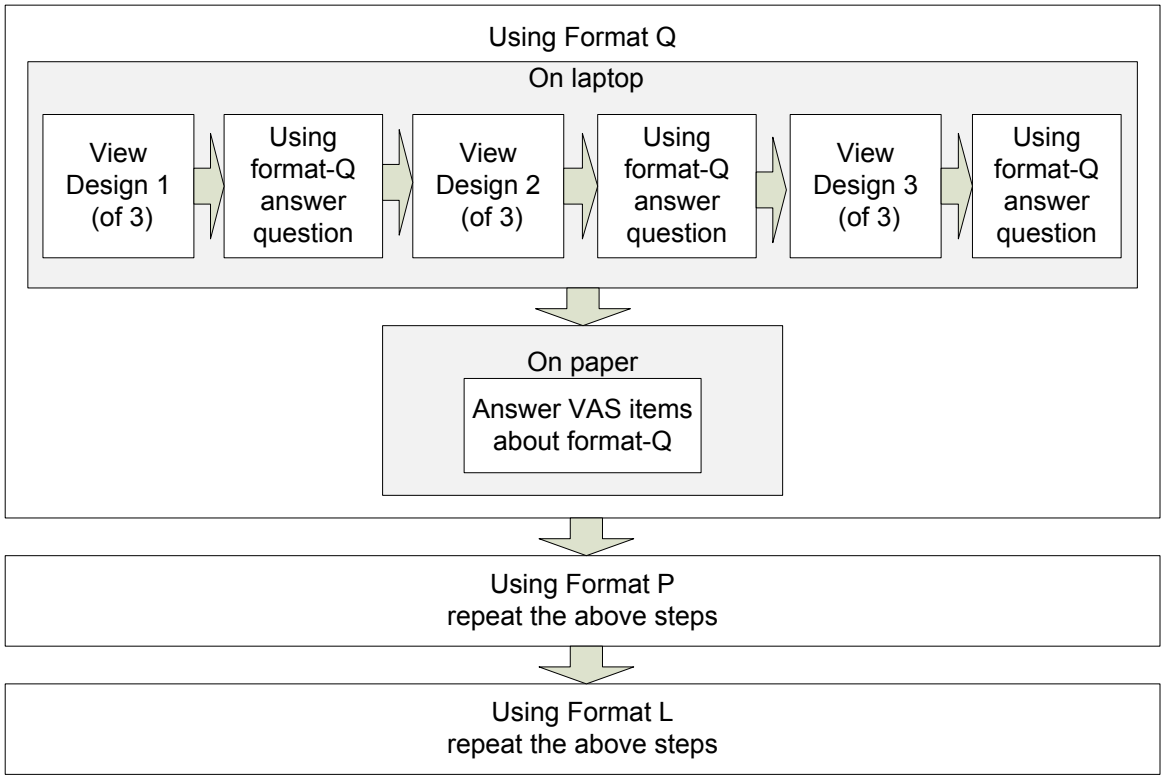


Figure 9.2 - Pilot evaluation experiment phase workflowfor the feedback participant task. The presentation order of formats was randomised (see page 188).

9.7 Feedback Side Results

9.7.1 The Conduct of the Tasks

The mean time on task for the 10 participants was 25 minutes (median: 23; SD: 8; max.: 44; min.: 17).

9.7.2 The Data

As data was collected on paper, steps were taken to ensure accuracy in recording and data entry. See Appendix B p.190.

The data from the VAS scale items was collated in spread sheets and analysed with SSPS.

The raw scores were found to be skewed towards the positive end of the scales; e.g. participants answering the utility item "How well were you able to express yourself?" tended to place their mark near the "Completely" end of the scale (positive) rather than the "Not at all" end (negative).

A log transformation was applied to all the scores to mitigate this skew (Equation 9.1) (Field, 2009).

$$f(x) = \log_{10}(x + 1) \tag{9.1}$$

The transformed distributions were tested for normality using the Kolmogorov-Smirnov (K-S) test (Field, 2009) and passed. (See Appendix B p.192 for details). It was inferred from this that parametric tests may be carried out on the distributions.

9.7.3 Means and Error Bar Charts

Figure 9.3, Figure 9.4 and Figure 9.5 show the mean log transformed VAS ratings for Utility, Freedom and Interest.

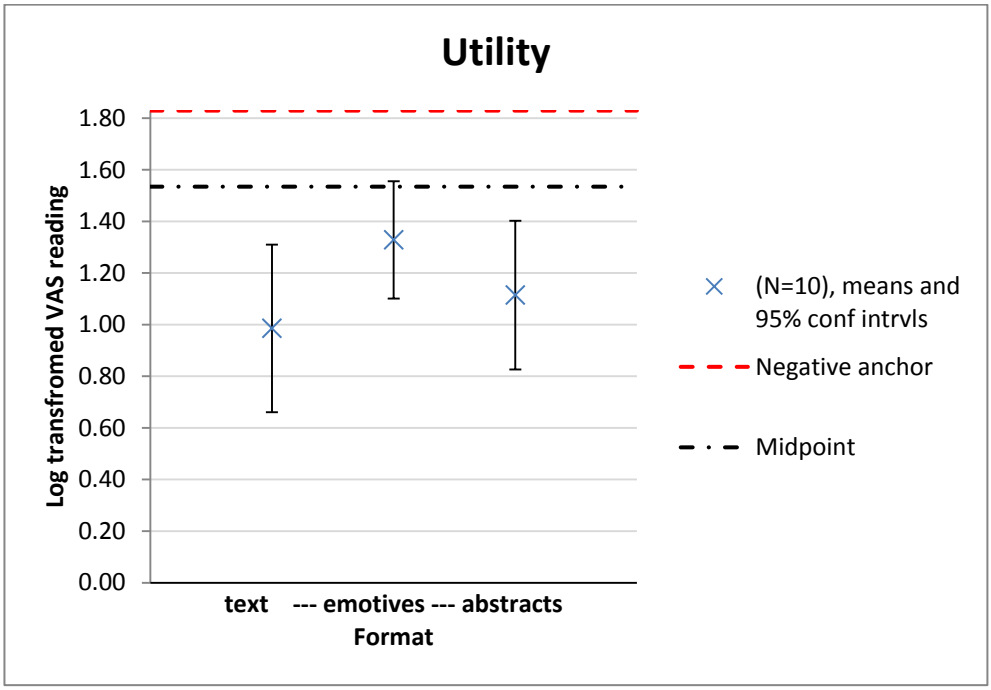


Figure 9.3 - Pilot Utility item transformed score means. Zero represents the positive anchor of the item scale. The scores representing the negative anchor and midpoints of the item scale are shown by the dashes and dot-dashed lines respectively. This is shown as a visual reminder that the y-axis shows log transformed scores.

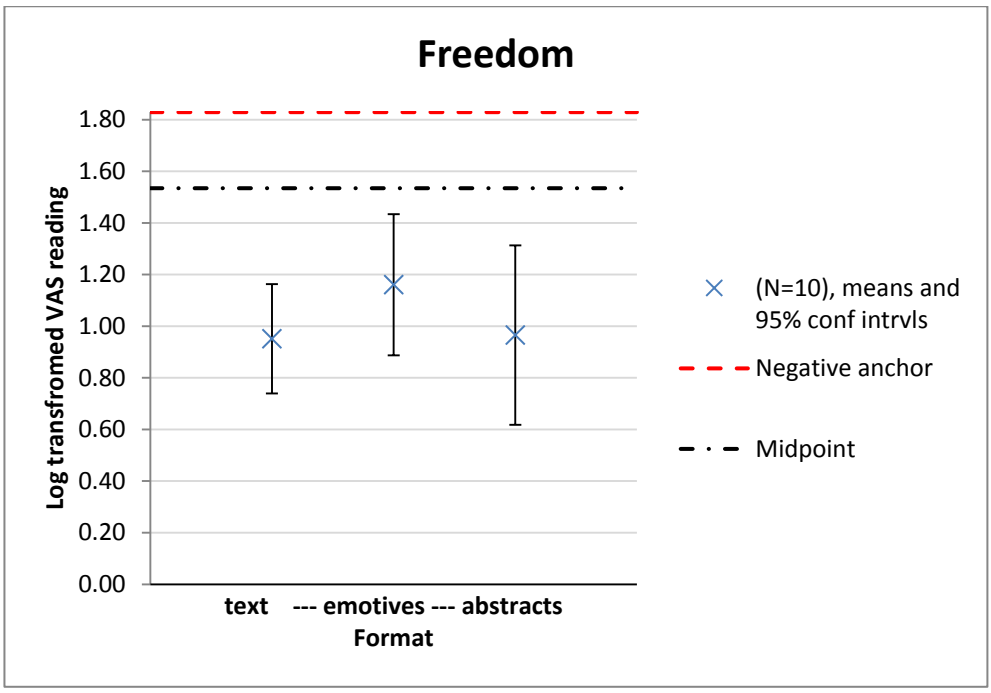


Figure 9.4 - Pilot Freedom item transformed score means. Zero represents the positive anchor of the item scale.

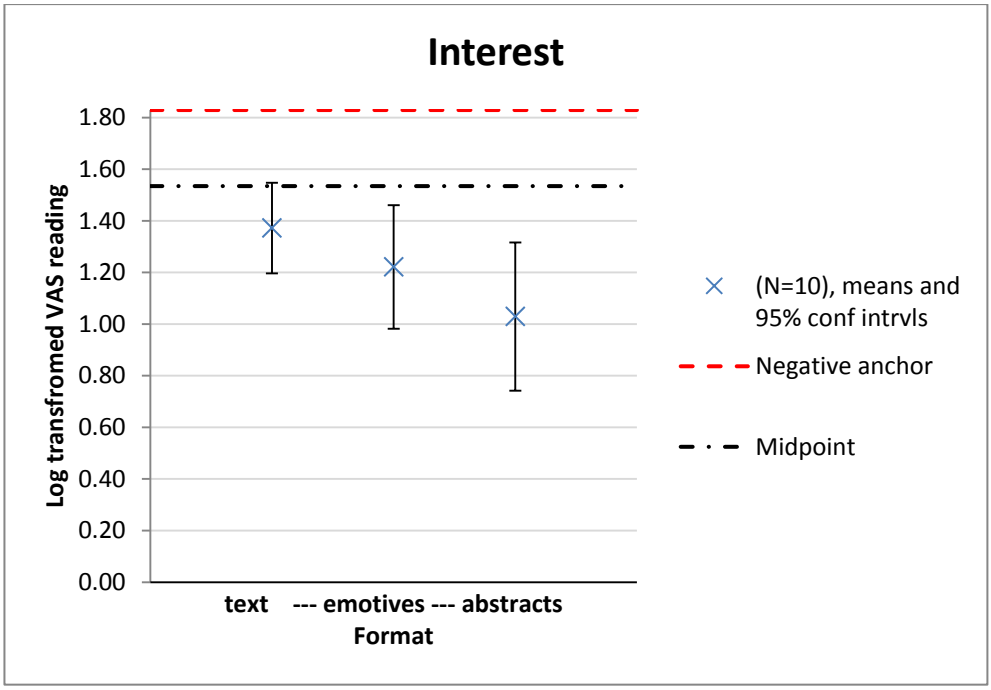


Figure 9.5 - Pilot Interest item transformed score means. Zero represents the positive anchor of the item scale.

9.7.4 ANOVA

Three one-way repeated-measures ANOVAs were carried out in SSPS, one for each of the three dependent variables: Utility, Freedom, and Interest. These are reported below:

Sphericity was not violated for any of the measures.

Utility

The results show that Utility as self-reported by participants was not significantly affected by the answer format, F(2)=2.48, p = 0.112 (i.e. not significant at the 0.05 probability threshold) .

Freedom

The results show that Freedom as self-reported by participants was not significantly affected by the answer format, F(2)=0.83, p = 0.453. (i.e. not significant at the 0.05 probability threshold) .

Interest

The results show that Interest as self-reported by participants was not significantly affected by the answer format, F(2)=2.48, p = 0.112. (i.e. not significant at the 0.05 probability threshold) .

As none of the measures showed a significant effect due to answer format the post hoc tests are not reported (Field, 2009).

9.8 Feedback Side Discussion

9.8.1 The Utility Measurement

Finding of no significant effect in the ANOVA due to answer format is interesting for Utility in that it indicates that participants feel they were able to express their emotional reaction to the designs equally well with text, emotive or abstract images. However there was an effect and while not significant at the 0.05 probability threshold the *p*-value is low (0.112). Looking at Figure 9.3 the mean for abstracts images was closer to that of text than was the mean for emotive images.

9.8.2 The Freedom Measurement

It might have been expected that participants would feel less inhibition when using images than using text to express their emotions. However, the finding of no significant

effect by the ANOVA of answer format on the freedom measurement does not support this nor does the chart, Figure 9.4.

9.8.3 The Interest Measurement

The chart, Figure 9.5, does indicate that participants tended to find the image formats, especially the abstract image format, more fun to use than text. However the ANOVA showed that the effect of answer format on Interest was not significant.

9.8.4 Evaluation of the VAS Items as a Whole and Individually

Here the VAS items are addressed, examined further where necessary, and decisions are made about whether these items used in the Pilot should be carried forward to the main experiment.

The VAS items

The participants all professed to understand the questions and the task so the format of using VAS items seems to work from the participants' point of view.

Interest

The Interest measurements although not producing a significant result in the ANOVA might show a significant effect with more statistical power from a larger sample size. This measure should be kept in the main study.

Utility

The Utility item was simply worded; it addresses an important question. This measure should be kept in the main study.

Freedom

The author suspected that participants might have found the Freedom item problematic a) due to the complexity of the question wording and b) due to confusion of the issues of Utility (ability to express) and Freedom (freedom to express). To seek confirmation of b) a correlation analysis was done on the Utility and Freedom raw untransformed scores (Table 9.7). This showed a strong correlation in particular for the text and

abstract images answer formats. This could be interpreted as confirmation that participants were conflating these two items. The ANOVA result indicates that any effect on this measure due to the answer format may be hard to demonstrate even with a larger sample size. Eliminating this item from the main study would a) decreasing the number of dependent variables from 3 to 2 and reduce the severity of any correction required for multiple comparisons and b) simplify and shorten the task for participants. The "freedom" issue could be addressed by a question in a post task questionnaire instead. These considerations informed the decision to remove this item from the main study.

Score pairing	**Pearson Coefficient *r***
Utility-Text vs. Freedom-Text	0.57 (large)
Utility-Emotives vs. Freedom-Emotives	0.14 (small)
Utility-Abstracts vs. Freedom-Abstracts	0.77 (large)

Table 9.7 - A correlation analysis of the Utility and Freedom raw scores. Each r value is accompanied in brackets by the descriptive category corresponding to that effect size (Field, 2009 p57).

9.9 Designer Side Interview Pilot

Although the pilot evaluation was primarily aimed at piloting the feedback participant VAS items the opportunity arose to pilot an interview format. One designer who had volunteered two of her designs for the feedback task agreed to attend an interview. The feedback was collated for each of her two designs separately. (However, there was only time in the interview to show the feedback for one of them).

9.9.1 Collating the Feedback Prior to the Interview

The manually collated feedback image selections were assembled into CSV files to mimic the format planned for the main study. These were then processed in MATLAB to produce the clustered summary definition files for input to the collage rendering web application. The text feedback was collated into randomly ordered lists in PDF format. There were no hurtful text comments.

9.9.2 Interview Script

A script of interview topics and questions was developed from the research questions (Kvale & Brinkmann, 2009 p.130) for the semi-structured interview. (See Appendix B p.192).

To address the issue of inspiration and in particular the research question, "*Are designers inspired by the visual feedback to make changes to their designs?*" two VAS items to be introduced within the interview were piloted. These were asked before and after the first feedback (the abstract image feedback summary) was shown to the designer participant. The questions were 1) "*How likely are you to make a change or changes to the design?*", and 2) "*At this moment how many design ideas do have in your mind?*". How they were presented as VAS items with anchors can be seen in Appendix B p.194.

9.9.3 Setting and Conditions

A simple web application was created to allow easy access to the design image and the 3 feedback formats during the interview. (See Appendix B p.194). The interview format required a means of a) displaying the designer's design image and the feedback formats for discussion and b) of allowing the designer to interact with the feedback formats.

Figure 9.6 - Interview setting. Items were displayed on a 24 inch monitor for discussion (controlled by researcher via a laptop and mouse). Designer participants interacted with feedback formats on an iPad3. This photo is a re-enactment with a fellow postgraduate student playing the part of a designer participant.

A laptop with additional 24 inch display was used to allow display of items for discussion and an iPad3 was used to allow the designer to interact with the feedback formats. The administrator via the laptop controlled the display of items on the monitor for discussion. The designer participant was prompted to tap particular buttons on the iPad to reveal specific feedback formats. When viewing the feedback formats the participant was encouraged to interact with them by tapping individual images on the visual summaries (the text list format could be scrolled). The setting for the interview was a room normally used for seminars or small-class lectures with Wi-Fi access (and cabled network access if Wi-Fi failed). (Figure 9.6 depicts the location that was used for half of the interviews in the main study. The other location used in the main study and that used for the pilot were similar in that they were seminar areas, with a window.)

9.9.4 Results and Discussion

The interview transcript, notes, and VAS items, were examined with a view to identifying what had not worked so that it could be left out for the main study interviews.

The paper VAS item readings were measured in the same way as for the pilot feedback participant task questionnaire and are shown in Table 9.8.

Question	Likelihood of making changes
Before	52.5
After	52.5
Question	Number of ideas
Before	71
After	70.5

Table 9.8 - Pilot interview VAS item readings The scale ranged from 0, the negative end, to 72mm the positive end. The readings were in mm. The "Before" reading was taken before viewing the abstract image summary and "After" taken after viewing the abstract summary.

PD1's design was a finished one but her response on the VAS item about how likely she was to make changes to the design indicated she felt more likely than not to make changes. She also indicated she had no shortage of design ideas. There was no change as a result of viewing the abstract feedback according to the VAS item responses. However by the time we reached the end of the interview it seems her thoughts on the abstract image summary had changed. After stating that she preferred the abstract image format over text and emotive images, when probed further she indicated that she

thought that the abstract format would show how the design was being received or provide inspiration for change:

> *Researcher – Do you think you would be taking inspiration from the abstract collage for future designs?*
>
> *PD1 – I think it does kind of influence me in terms of a creative way and kind of makes me feel how people would want to see the design change or how they are interpreting it. It actually does really help to see how, what emotion of my design evokes through images like that.*

These VAS items, on the issue of inspiration, while consuming valuable interview time had not contributed reliable information. Perhaps they were valid but had been asked in too close proximity for the feedback to sink in? Perhaps because her individual design was a finished one in this case she had not been inspired to change while later she had considered use of the feedback generally and in future? Unfortunately it had not been possible to follow this up in the interview (time had been running short). These items broke the flow in the interview and the issue was able to be addressed through considered discussion in the interview such as that quoted above. Therefore those VAS items should not be used in the main study.

Additionally: in both the items the before and after measurements were almost exactly the same, to the within a millimetre. This is confirmation of the reliability with which respondents can gauge where to place their mark for a given self report opinion.

The question asking for a single word to describe each format (asked after viewing and discussing) did not seem useful. She described the abstract summary as “Bold”, the emotive summary as “Emotion” and the text list as “Literal”. Given that the author had termed the emotive image summary as “Emotive image collage” and that text can literally be described as “literal” these three questions had not yielded value for interview time. They also interrupted the conversation flow where more might be gained from following up other answers and letting participants freely express views rather than artificially tying them down to a single word. For these reasons those questions should not be used in the main study.

The other questions had all provoked interesting answers, such as the one quoted above, and so it was decided to retain the rest of the pilot script but be more prepared to pursue some answers with follow-up questions.

The decisions, from the above discussion, on how to proceed with the interviews in the main evaluation are summarised below in 9.10.3.

9.10 Conclusion

The aims of the evaluation were established; the evaluation is to be considered as having a feedback side and a designer side; and, a format for the main evaluation was chosen. On the feedback side: feedback participants will view stimuli (each being a design and a stimulus question), respond in three answer formats (generating the feedback), and give VAS judgements about the answer formats (generating interval data). The designer side will use semi-structured interviews in which designer participants view feedback (generating qualitative data). It was decided to conduct a pilot to confirm the viability of, and to rehearse, some of the methods.

This section continues with an overview of the pilot study. Then the decisions on how to proceed with the feedback side from section 9.8.4 are summarised. Lastly, the designer side decisions from the pilot are summarised.

9.10.1 Overview of the Pilot Study

Both the feedback task pilot and the interview pilot were helpful. The feedback pilot showed that feedback participants were comfortable with using the VAS items and that two of the items had produced data with a good prospect of being analysed successfully in the main study. It also showed that the feedback participants were highly positive about using the formats (i.e. the raw VAS scores were skewed towards the positive anchors). The interview pilot showed that the rather unconventional idea of using VAS items within the interview did not work well, but other questions and the setting arrangements were good. It had also provided useful practice prior to the main study.

9.10.2 Feedback Side Task Decisions

Summary of decisions for proceeding to the main evaluation:

1. Keep the Utility and Interest VAS items.
2. Discard the Freedom VAS item. Address the issue of freedom of expression in a post-task survey.

9.10.3 Designer Side Interview Decisions

Summary of decisions for proceeding to the main evaluation:

1. Keep the setting and conditions. These worked well.
2. Keep most of the interview script (with exceptions detailed in 9.10.1) but be more prepared to probe and follow up some answers.
3. Discard the VAS items about inspiration used during the interview.
4. Discard the question asking for one word to describe each feedback format.

Chapter 10

Main Evaluation Study

One goal of this thesis was "*to develop the means to implement this method of crowdsourced visual feedback sufficiently to allow its evaluation"*. The means were developed in Chapters 2 to 8. Chapter 9 began the evaluation by setting out the study design, piloting that study, and concluding with the adjustments to the study design ready for this main evaluation. In that chapter Table 9.3 established *Evaluation Research Questions* and this led to considering the evaluation as having two sides: the *Feedback side* and the *Designer side*.

For the *Feedback side* Chapter 9 concluded that the main evaluation should consist of feedback participants doing two activities:

1. A feedback task to
    a) Gather design feedback in the three formats (Text, Abstracts, and Emotives), and
    b) Measure the Utility and Interest of the formats with VAS items;
2. A post-task survey to further probe their views about the formats.

For the *Designer side* it was established that this should consist of these steps:

1. Designer participants provide their designs for feedback.
2. Feedback on the designs is gathered during the feedback task.
3. Feedback is collated and summarised.
4. Designer participants attend semi-structured interviews during which they view the feedback and describe what they think of it.

The remainder of this chapter describes the main evaluation thus:

The *feedback side* is described in Sections 10.1 to 10.3. Section 10.1 details two aspects affecting the methods used in the feedback task. The task methods and the post-task survey are set out in Section 10.2. Lastly on the *Feedback side,* Section 10.3 reports the results from the feedback task and post-task survey. It integrates those results and draws

some conclusions including conclusions about a division in the feedback participant group.

The *designer side* is dealt with in Section 10.4, detailing the methods; reporting and discussing the results.

Section 10.5, Discussion and Conclusions for the whole of the main evaluation, summarises the results from the *feedback side*, brings in feedback side results from the pilot study for comparison, and discusses the *designer side*. It goes on to discuss the two image types (abstract and emotive), revisit the *Evaluation Research Questions*, and finally the section reports how designer participants viewed the possibility of a new web service providing visual feedback.

Appendix C is the appendix associated with this chapter.

Published work

The studies described in this chapter and in Chapter 9 feature in published work: Robb et al (2015a) and Robb et al (2015b).

10.1 Design of the Main Study Feedback Task

10.1.1 A Tension in the Design

There were two conflicting imperatives in the design of the feedback task for the main study:

- The need to maximise the number of feedback images per design so as to produce enough images to summarise, and enough items of text feedback to make the body of feedback seem substantial to the designer participants.
- The need to minimise the number of items each feedback participant had to answer so as to keep the time on task acceptable and avoid fatigue.

10.1.2 Change to the Workflow from the Pilot

It was decided to have feedback participants provide VAS judgements about each answer format after viewing each design, i.e. repeatedly, during the experiment phase

rather than just once after several uses of a given answer format. This repetition would have the benefit of generating more readings for each measure and thus should decrease the experimental error (as each reading for a measure would be the median of several readings) and increase the signal to noise ratio in the final VAS readings. There was a downside risk of fatigue for the participants through repetition.

10.2 Feedback Task

10.2.1 Interface and Recording Method

Having decided on a task format which required an interface to gather and store data from participants (see 9.2.1), a web interface was implemented in PHP, JavaScript (using jQuery), and MySQL. It served the stimuli and recorded feedback responses and VAS item judgements in a database. The stimuli were served according to stimuli packets generated in MATLAB and stored ready in a database. Screens from the interface can be found in Appendix C (p.197). The VAS readings ranged from 0 to 383 (the number of pixels used to display the scale in the web application (Reips & Funke, 2008)).

10.2.2 Training Phase Work Flow

As with the pilot, there was a training phase to allow participants to experience all three answer formats and both of the VAS items. The training phase consisted of one unit of work illustrated in Figure 10.1. This workflow differed from the pilot (Figure 9.1) in that this time the VAS items were completed immediately after each use of an answer format. The order of presentation of the formats was randomised for each participant. As in the pilot the formats were labelled with letters to avoid any preconception of precedence or concepts in the participants ("Q" for text, "L" for abstract images, and "P" for emotive images).

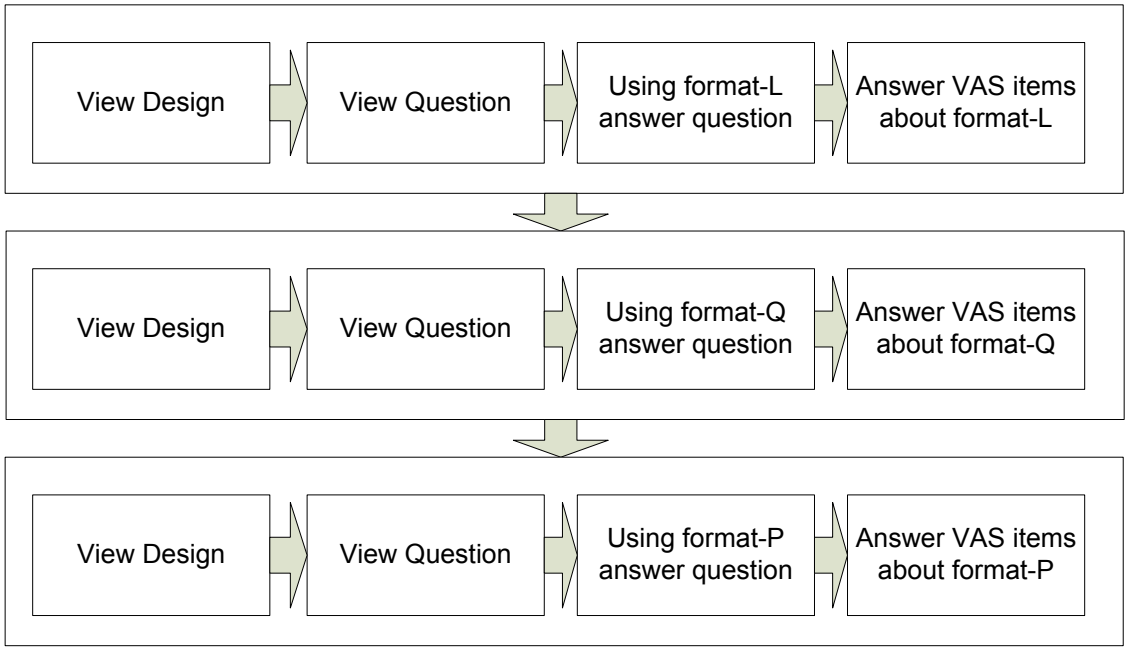


Figure 10.1 - Feedback task work flow for one unit of work. The same design and question are viewed three times. Each time the participant used a different format to provide their feedback response and then gave their opinion about that format using the VAS items. The order of formats e.g. L-Q-P was randomised for each participant.

10.2.3 Experiment Phase Workflow

The experiment phase workflow was the same as for the training phase but consisted of five units (Figure 10.2). The whole workflow presented one design during the training phase and five during the experiment phase. Thus a total of six designs were viewed and six sets of three feedback responses were gathered per participant. Five sets of VAS measurements were gathered for each of the three response formats during the experiment phase (the training set being discarded prior to analysis).

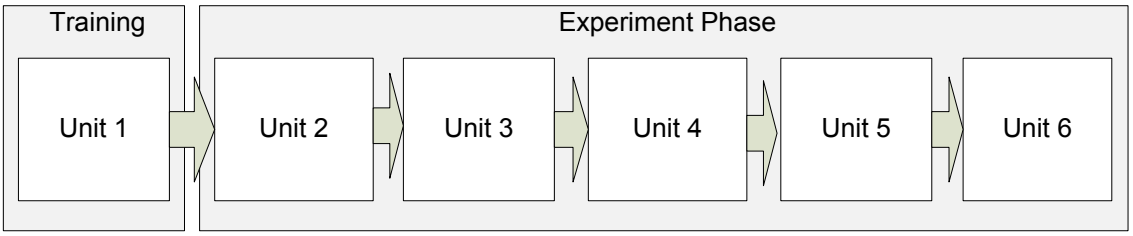


Figure 10.2 - Feedback task overall workflow. Each unit presented a different design and had workflow as shown in Figure 10.1.

10.2.4 Post-Task Survey

The pilot study feedback task sessions, having been administered on an individual and personal basis, had all finished with an additional informal debrief to establish if the participant had understood the task and the questions, to give the participant the opportunity to comment, and to thank them. There would be no such individual opportunity for the main study feedback task as the sessions were being conducted concurrently en-masse. Instead, feedback participants were asked to complete a short web-based survey after completing the task. The purpose of the questions fell into these categories: 1) participant ID fields to allow the survey answers of each participant to be matched anonymously to their task data; 2) establishing whether or not the participant had understood what they were doing in the task; 3) seeking opinions about the visual feedback formats; 4) an opportunity for open-ended comment; and 5) to ask participants to report specifically on the issue of "freedom of expression" because, from the pilot, it had been decided to discard the VAS item measuring this during the task (see 9.10.2).

Details of the survey are in Appendix C (p.198). Results from the survey are included in the discussion of the results below.

10.3 Feedback Side Results

10.3.1 The Conduct of the Tasks

The participants (see 9.2.2) assembled in a lecture theatre and, after reading and signing the consent form (Appendix C p.196) and listening to a brief explanation, they retired to two computer rooms to do the task and post-task survey. One participant took part from home. The mean time on task for the 32 participants was 19 minutes (median: 18; SD: 5.8; max.: 35; min.: 10).

All except one participant completed the post-task survey. One comment in the post-task survey stated that the task was too repetitive. The number of items that each feedback participant was served had been one of the issues considered in 10.1.1 and had been of concern. Therefore, mean readings for each of the sequence of five experiment phase readings across the 32 participants were examined and are shown in the chart in Figure 10.3. A reading of zero equates to the positive anchor of a given VAS item e.g. for the interest item, "Very Much Fun". A reading of 383 equates to the negative end

e.g. “Very Much Boring”. An upward trend might have indicated that fatigue had negatively affected the participants’ judgements. For all six measurements (three formats by two readings each) the mean score rises slightly (i.e. becomes more negative) from reading 2 to reading 4, but equally, all except one drop from reading 4 to 5. The Utility_Text mean scores appear to vary the most across the sequence of readings. However the differences between first and last readings are not large relative to the range of the scale (0 to 383). It was concluded that participant fatigue had neither a consistent nor a marked effect on the mean scores over the five readings for the six measurements on the whole.

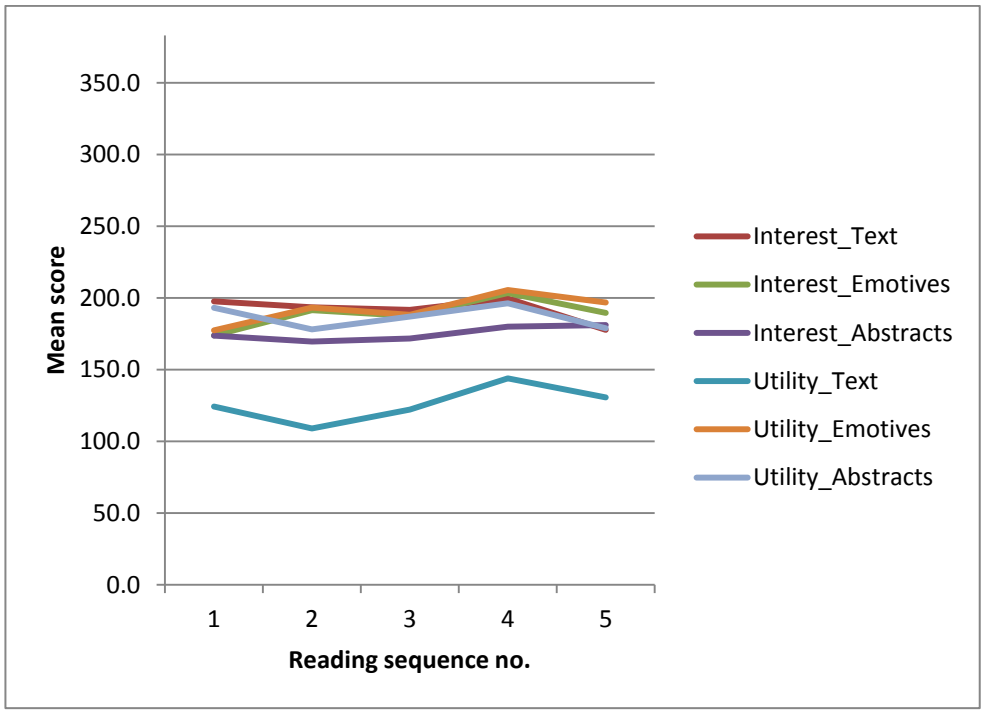


Figure 10.3 - Checking for participant fatigue affecting the results: The mean VAS item scores across all 32 participants for the sequence of five experiment phase readings. 0 marks the positive anchor, 383 marks the negative anchor. Thus an upward trend would have indicated that scores (representing judgements about the feedback formats) were becoming more negative over time and fatigue might be the cause, which did not appear to be the case.

10.3.2 The Data

The data from the VAS scale items was gathered by running queries on the recording database. The training phase readings were set aside. For each participant there were six measurements (three formats by two VAS readings each) for each of the five experiment phase units (Figure 10.2). Thus for each participant there were six sets of five readings consisting of integer values from 0 to 383. The median of the five readings

was taken to represent a given participant's response. Therefore, following this initial processing, for each of 32 participants there were six median readings: *Utility_Text, Utility_Emotives, Utility_Abstracts, Interest_Text, Interest_Emotives,* and *Interest_Abstracts*. (See Appendix C p.200 for these detailed results).

The distributions, unlike in the pilot, did not require to be log transformed. (In the pilot the scores were skewed towards the positive end of the scale.) All six score distributions were tested for normality using the K-S test (Field 2009) and passed. We inferred from this that parametric tests may be carried out on the distributions.

10.3.3 Means and Error Bar Charts From the VAS Items

Figure 10.4 and Figure 10.5 show the means and 95% confidence intervals for text, emotive image, and abstract image formats for Utility and Interest respectively.

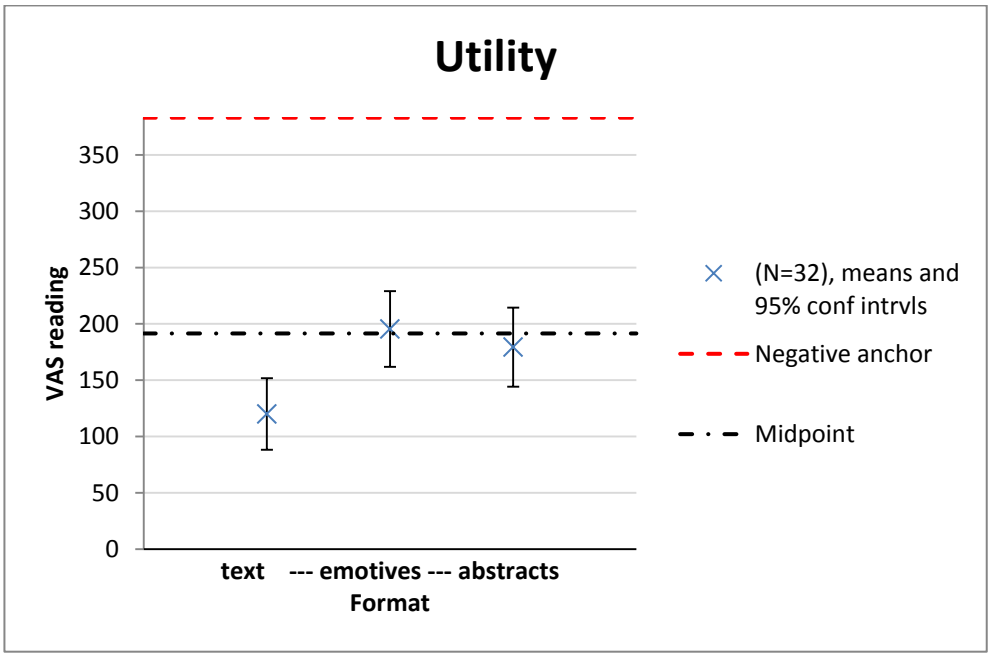


Figure 10.4 - Main study, Utility item score means. Zero represents the positive anchor of the VAS item i.e. "Completely" able to express an answer using that format. 383 represents the negative anchor i.e. "Not at all" able to express an answer. The midpoint of the item scale (not explicitly marked for participants) is shown by a dot-dashed line.

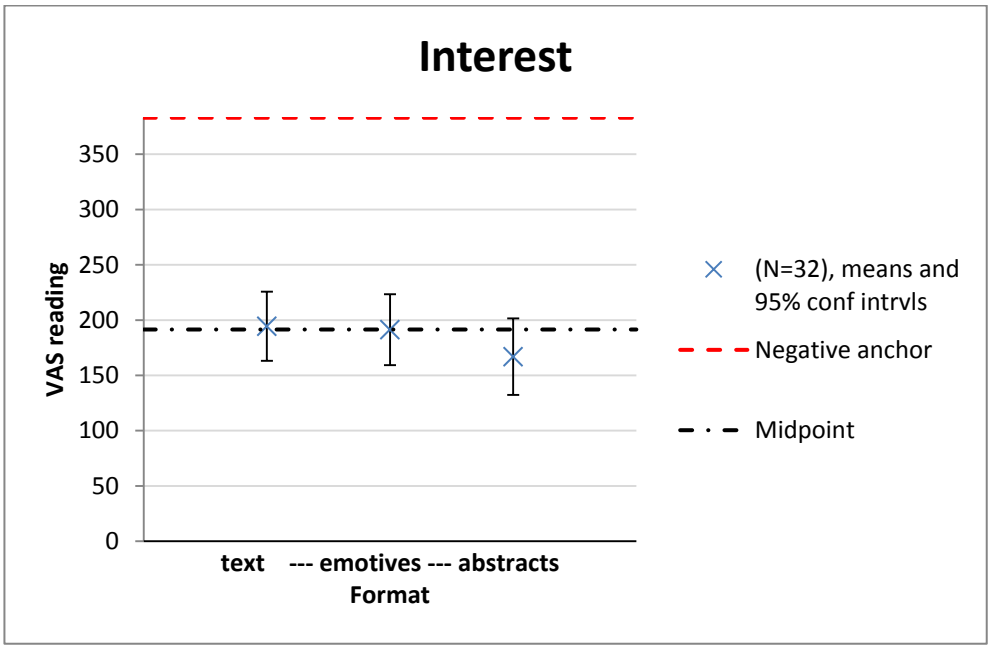


Figure 10.5 - Main study, Interest item score means. Zero represents the positive anchor of the VAS item i.e. "Very much fun" using that format. 383 represents the negative anchor i.e. "Very much boring".

10.3.4 ANOVA on Whole Feedback Group

The reader is reminded to bear in mind that *lower is better* on these two VAS items; i.e. a low reading represents one closer to the positive anchor on the item, a low score is better than a high score in terms of either Utility (self-reported ability to express an answer using a given format) or Interest (self-reported level of interest when using a given format).

A one-way repeated-measures ANOVA was carried out in SSPS for each of the two dependent variables: Utility, and Interest. These are reported below:

The ANOVA for Utility

Mauchly's test indicated that the assumption of sphericity had been violated, $\chi^2(2)=9.57$, $p<0.001$, therefore degrees of freedom were corrected using the Greenhouse-Geisser correction ($\varepsilon = 0.78$). The ANOVA showed that the participants ability to express themselves, as measured by the Utility self report VAS item, was significantly affected by the answer format, $F(1.57, 48.70) = 12.60$ $p<0.001$. Post hoc tests using the Bonferroni correction showed that Utility was slightly better for abstract

images (M=179.3) than for emotive images (M=195.5) but this was not statistically significant (p = 0.45). However, Utility for text (M=120.1) was significantly better than for emotive images (p=0.001) and abstract images (p=0.006).

This leads us to conclude that, the group of feedback participants as a whole, reported being better able to express their answer using text compared to using the image formats.

The ANOVA for Interest

Mauchly's test indicated that the assumption of sphericity had been violated, $\chi^2(2)=18.58$, $p<0.001$, therefore degrees of freedom were corrected using a Greenhouse-Geisser correction ($\varepsilon = 0.68$). The results show that participants' level of interest, as measured by the Interest self report VAS item, was not significantly affected by the answer format, $F(1.37, 42.42)=1.93$, $p>0.05$ (=0.17).

As the Interest measure showed no significant effect due to answer format, post hoc tests are not reported (Field, 2009).

10.3.5 Feedback Participant Preferences

The feedback participants were asked in the post-task survey to rank the three answer formats, text, abstracts and emotives, by overall preference (forced ranking). 31 of the 32 feedback participants responded and a quantitative analysis of their preferences is reported below. Table 10.1 shows the frequencies with which each ranking was awarded. The mean rankings are calculated by giving the frequency of each ranking a weight equivalent to its rank and dividing by the total number of responses (e.g. for Abstracts its mean ranking of 1.81 = ((15x1)+(7x2)+(9x3))/31). Note that a low value means a better mean ranking i.e. 1.0 would have been the best possible mean ranking.

Rank / Format	1	2	3	Responses	Mean ranking (1 is best; 3 is worst)
Abstracts	15	7	9	31	1.81
Emotives	5	14	12	31	2.23
Text	11	10	10	31	1.97
Total	31	31	31		

Table 10.1 - The overall preference ranking frequencies of the three formats by the 31 feedback participants who responded. Abstracts and Emotives were the two image formats.

Figure 10.6 compares text with images (either abstract or emotive) by showing the frequency with which participants ranked text as their first preference against those ranking one of the image formats as their first preference (i.e. 15 for abstract plus 5 for emotive totals 20 feedback participants who ranked one of the image formats as their most preferred answer format).

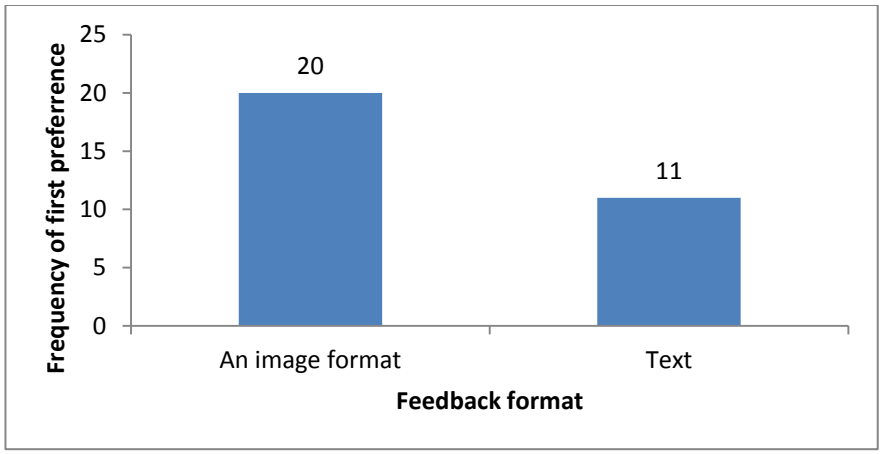


Figure 10.6 - Chart showing the frequency with which an image format and text were ranked as first preference by the 31 feedback participants who responded.

Cognitive styles theory (see 2.5) would predict that, taking into account the visual-verbal dimension, some people are more visual and others are more verbal. This appears to be borne out in these results in that 11 of the feedback participants preferred to respond using text while 20 preferred responding with images.

10.3.6 Considering Feedback Participants as Two Groups

The feedback format preferences suggested that there might be two populations represented within the feedback participant group. This prompted further analysis of the data from the feedback task. The VAS item median readings were split into two groups 1) from the 11 participants who stated text as their first preference for feedback format (text-likers) and 2) from those 20 whose first preference was an image format (image-likers). (See Appendix C p.200.) The readings from the participant who did not complete the survey were set aside. The means of the readings from the two groups are compared in Figure 10.7. These charts show an interesting picture. It is clear that the text-likers and image-likers are behaving differently. There is little difference between their perception of the text format. However, there is a marked difference between their

perceptions of the image formats, with the image-likers finding the image formats both more fun and more useful for answering than did the text-likers.

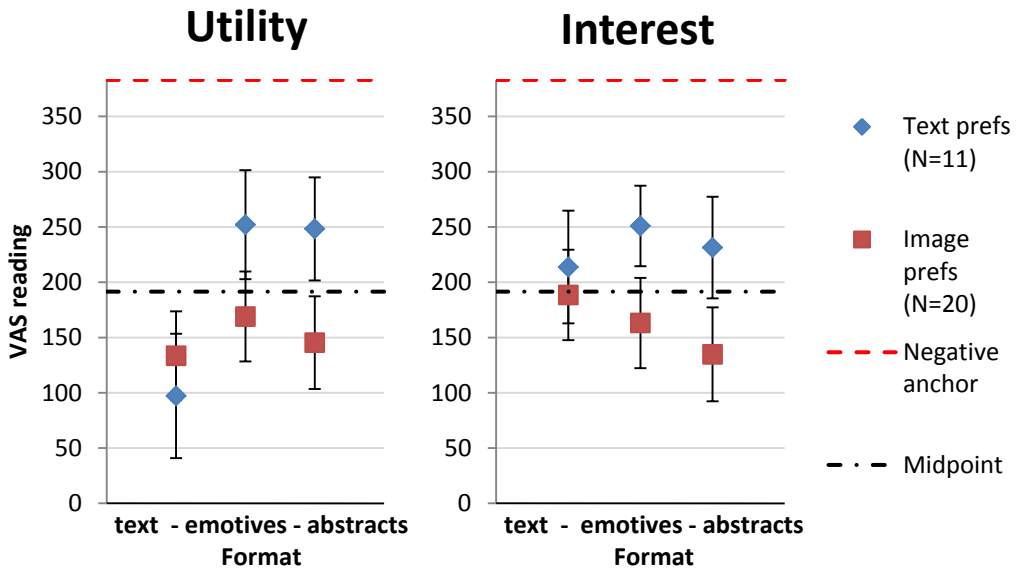


Figure 10.7 - Main study, Utility and Interest item means, by groups, with 95% confidence limits (error bars).

As the ANOVA on the 32 participants had shown that they had found text more useful for answering than images, another ANOVA was done on just the Utility measure for the image-likers and this is reported below:

Mauchly's test indicated that the assumption of sphericity had not been violated, $\chi 2(2) = 1.014$, $p = 0.602$. A repeated measures ANOVA showed that the 20 image-likers' ability to express themselves, as measured by the Utility self report VAS item, was significantly affected by the answer format, $F(2, 38) = 3.556$, $p = 0.038$. However, post hoc tests using the Bonferroni correction showed that, despite there being differences between the mean Utility for text, emotives, and abstracts (M= 133.5, M=169.0, and M=145.4 respectively) these differences were not statistically significant. (Text vs. emotives, $p = 0.085$; text vs. abstracts, $p = 1.000$; emotives vs. abstracts, $p = 0.276$).

The ANOVA result shows that, for image-likers the difference between their perception of the Utility of text and the Utility of abstract images was *not* statistically significant. Whereas, for the 32 participants as a whole, the difference between Utility for text and Utility for images *was* statistically significant.

Thus it is concluded that the feedback participants consist of two groups: one preferring text and the other preferring images. The image-likers showed no statistical difference

in their judgement of the Utility of text compared to images (particularly abstract images), implying that they felt able to describe their emotions using images (whereas the text-likers judged Utility of text to be better than images). On the measure of Interest, although a statistical difference has not been shown, the chart for interest in Figure 10.7 shows that, for image-likers, the Interest_text mean (M=188.6) straddles the midpoint of the Interest scale, whereas the Interest_abstracts mean (M=134.8) and confidence limits lie to the "fun" side of the scale. In addition, in that chart, image likers do appear to have judged the image formats as more fun to use than the text-likers did.

10.3.7 The Freedom Theme

The issue of feedback participants feeling more or less inhibited (from ERQ 4, Table 9.3) was addressed by a question in the post-task survey. *Nvivo* text analysis software was used to analyse the survey responses using a similar method to that used for the designer interview analysis (10.4.4). The detailed results from this question are in Appendix C (p.201). Themes arising from the responses and the frequency with which they were expressed are summarised in Table 10.2. The majority of the responses were off-topic and those off-topic themes are discussed later.

Theme: Sub-theme	**Number of responses**
Not holding back irrespective of format	2
Holding back: When using Text	5
Holding back: When using Emotive images	1
Abstract images were not hurting feelings	1

Table 10.2 - Summary of themes from the post-task survey concerning ERQ4 (Table 9.3).

Of those feedback participants who directly addressed the issue of whether or not they held back to spare the feelings of the designers, the majority held back when using text. With the image formats, only one participant stated that they held back using emotive images, but none held back with the abstract images; indeed one participant stated positively that abstract images would not hurt designers' feelings.

The preponderance of off-topic (but interesting) responses suggests either that the participants found it difficult to self-report on this issue, or that the question wording had not been effective at probing the issue.

10.3.8 Other Themes from the Post-Task Survey

Several other themes arose from the grounded theory analysis of the post-task survey. Figure 10.8 shows the themes and a quantitative analysis of the frequency with which participants expressed them.

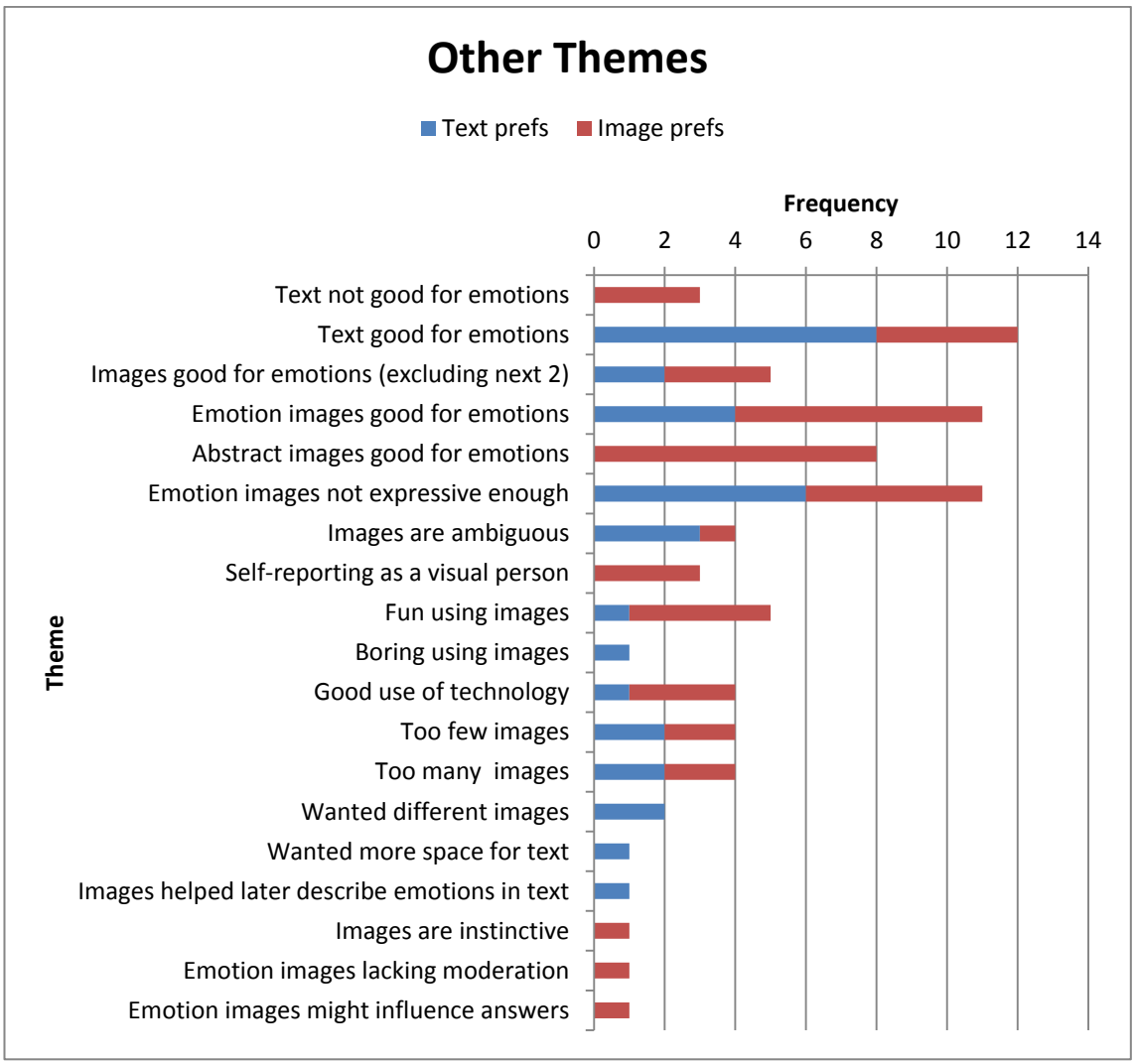


Figure 10.8 - Other themes from the post-task survey and the frequency with which participants (both those who preferred text and those who preferred image formats) expressed them.

Notable observations on Figure 10.8 are:

- Although 12 participants (including 4 image-likers) opined that text is good for expressing emotions, 24 (including 6 text-likers) stated images (abstract or emotive) were good for expressing emotions. Some of the participants think both images and text are good for expressing emotions.

- 11 participants (including 5 image-likers) were dissatisfied with the emotive expressiveness of the Emotive204 image set with one participant specifically stating there were not enough moderate emotion images.
- 3 participants felt it relevant to volunteer that they were "visual" people, indicating that they already feel themselves predisposed to images as a medium. (There was no mention or hint of the visual-verbal cognitive style dimension in the survey questions, in any of the materials or in task instructions.)

10.4 Designer Side Interviews

10.4.1 Collating the Feedback Prior to the Interviews

To maximise the body of feedback for each designer it was decided to include the text and image feedback from the training phase of the feedback task (rather than discard it). See Appendix C p.202 for the considerations involved in this.

The feedback text and image selections for each designer were collated by running queries on the feedback task database. These produced two image selection lists (ISL) as CSV files for each designer. The abstract and emotive ISL files were then processed in MATLAB (along with their associated 3D non-metric MDS coordinates files and similarity matrix or emotion tag vectors files) to produce the summary cluster definition files for input to the collage rendering web application. (See Appendix C p.202 for details). The text feedback was collated into randomly ordered lists in PDF format.

With each of the 32 feedback participants responding to half of the 12 designs, approximately 16 sets of responses we were collected for each design. Therefore a typical set of responses for a design consisted of 16 text responses, 48 abstract image selections, and 48 emotive image selections.

10.4.2 Interview Script

The semi-structured interview script was adapted from that developed for the pilot (9.9.2) in line with the conclusions in 9.10.3. (See Appendix C p.203).

10.4.3 Setting and Conditions

The setting and conditions were the same as for the pilot (9.9.3). The appearance and interactivity of the visual summaries and text lists was just like those from the pilot shown in Appendix B pp.194 - 196 with the exception that the text lists were longer, consisting of around 16 items.

The semi-structured interviews were approximately 45 minutes in length. Each started with a 15 minute warm-up consisting of a walk-through of the two image sets (abstract and emotive), how they were constructed and how selections from them can be summarised as a smaller number of representative images. The designers were asked to talk about how they used images in the design process and about their designs so as to establish the development stages of the designs. During the rest of the interview the three forms of feedback were revealed to the designer in a random order (recorded in Appendix C p.203) and their views were probed in line with the script. Additional questions followed up points raised by the designer participants.

10.4.4 Analysis Method

Audio recordings were made and transcribed. By following a grounded theory approach, using open coding (Corbin & Strauss, 2008), themes were identified. *Nvivo* text analysis software was used to facilitate this (Bringer et al., 2006) (Silverman, 2010). Some quantitative analysis was also carried out.

10.4.5 Results and Discussion

Themes from the interviews

Detailed descriptions of the themes with supporting quotes from the interviews are in Appendix C pp. 205-210. The themes are summarised below in Table 10.3.

Theme No	Description	Summary of theme
1	Interpreting the feedback	Participants developed their interpretation of messages from the feedback even when initially they perceived ambiguity.
2	Inspiration to make changes	The visual feedback inspired changes. Specific changes motivated in those whose designs were prototypes and ideas for the future described by those whose designs were more developed. A quantitative analysis of inspiration following first feedback showed the following: Text: 0/3; Abstracts: 2/4: Emotives: 3/4; i.e. 5/8 were inspired to form of change by image feedback while comparably none were inspired by text.
3	Abstract image summaries as mood boards	Abstract image summaries can act as "reverse-engineered" mood boards: as positive confirmation that the intended mood was received; and in the negative showing that the wrong mood was received thus motivating changes such as to colours and textures in the next design iteration.
4	Negative feedback	Merited subdivision. See 4.1-4.3
4.1	Perception of negative feedback across formats	Abstract image feedback is seen as non-threatening. Negative feedback such as "boredom" was read in emotive images.
4.2	A tendency to focus on negative feedback	Participants sought out negative feedback, skipping straight to negative text comments and focussing in on negative emotive images despite the majority (70%) of feedback being positive.
4.3	The impact of negative text compared to negative emotive images	There was disagreement between designers: e.g. one stated negative feedback in text was more impactful than emotive images, while another stated the opposite.
5	Effectiveness at finding out how people felt	Some designers thought emotive images had enabled feedback participants to communicate emotions more effectively than text
6	A service offering the visual feedback	Merited subdivision. See 6.1-6.2
6.1	Would designers use the visual feedback service?	11 out of the 12 participants wished to use such a service. Designers valued the visual feedback formats and wished to continue receiving visual feedback.
6.2	Present prototypes and refine through cycles of visual feedback	The participants were unanimous that this is how they wished to use the service.

Table 10.3 - Summary of themes from the interviews.

Designer participant format preferences

One part of the interview format involved asking participants to decide which feedback format they most preferred and least preferred and ask them to elaborate on the reasons for their preferences. Table 10.4 summarises these preferences. (For detailed results see Appendix C p.210).

Rank / Format	1	2	3	Responses	Mean ranking (1 is best; 3 is worst)
Abstracts	5	3	4	12	1.92
Emotives	2	5	5	12	2.25
Text	5	4	3	12	1.83
Total	12	12	12		

Table 10.4 - The overall preference ranking frequencies of the three formats by the 12 designer participants.

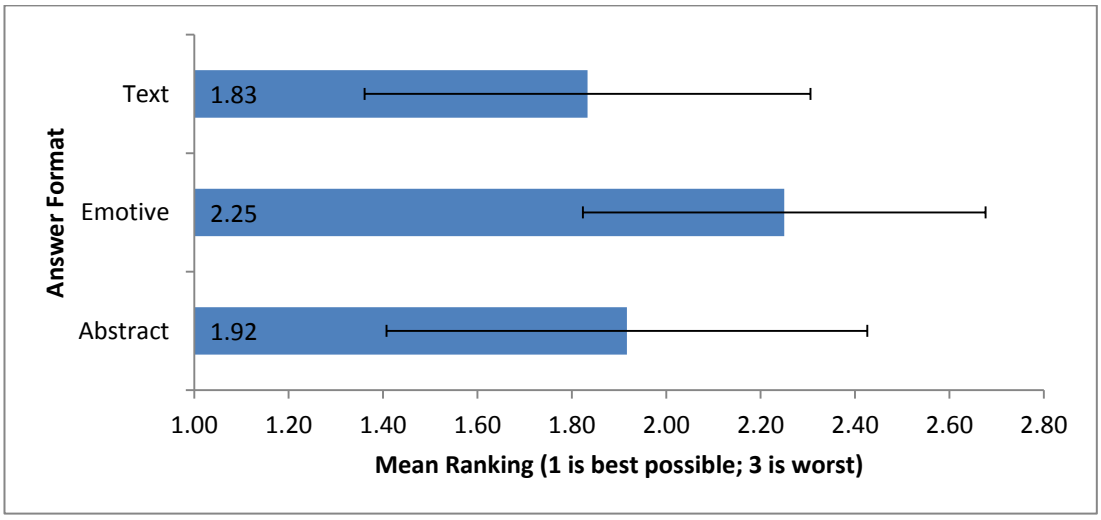


Figure 10.9 - Designer participant format preference mean rankings with 95% confidence limits.

Figure 10.9 shows the mean format rankings with 95% confidence limits. Thus it can be seen that the numerical results from this line of questioning from the interviews do not show any statistically significant difference between the formats based on the mean preference rankings.

The quantitative analysis of text versus image format first preferences shows 5 designer participants ranked text as their first preference versus 7 (5 abstract + 2 emotive) ranked an image format first. Thus it might at first seem that the designer participants may be splitting along the visual-verbal cognitive dimension. However, although one designer participant does self-report as a "visual person", the reasons participants gave for their chosen preferences appear to include much considered motivation. Table 10.5, below, summarises the designer participant format preference ranking themes (or FPR-themes). For the detailed analysis see Appendix C pp.211-213.

FPR-Theme No.	Themes from the reasons for ranking a given format first	Partic-ipant
Text		
T1	There is detail in the text.	D11
T2	Easier to get the meaning from text.	D11, D12, D5
T3	Text is "honest" (Less able to avoid the issue when you read it). (See A4 and E3)	D2
T4	Images can be ambiguous.	D12 ,D4, D5
Abstracts		
A1	Self reporting as a visual person.	D10
A2	Abstracts show how the design is perceived and understood.	D10, D3, D7
A3	Text is just less interesting. There is depth to the images.	D1
A4	Abstracts are open to interpretation allowing the reader to avoid or overlook negative feedback (compared to emotives and text). (See T3 and E3).	D3
A5	Abstracts are less negative than emotives.	D3, D6
A6	Abstracts are more understandable than emotives. (See E1)	D7
Emotives		
E1	Emotives give more meaning. (See A6)	D8
E2	Text is too conventional.	D8
E3	Emotives are open to interpretation allowing sensitive viewers to avoid negative feedback. (See T3, and A4)	D8
E4	Emotives can convey negative feedback.	D8
E5	Emotive images made the crowd reflect more on their emotions.	D9
E6	The emotives gave a different perspective on the design.	D9

Table 10.5 - Summary of the themes from the designer participants' reasons for ranking a given format first. The themes are attributed to participants giving each theme a quantitative weight.

10.5 Discussion and Conclusions

Before discussing the results it should first be pointed out that the nature of the sample for our main evaluation means that generalising from our findings should only be done cautiously. The main study feedback participants may not be representative of the general population; as students in a contextual studies course it is possible they could hold some non-typical attitudes about design communication and imagery; also they were predominantly female (although cognitive styles are said to be independent of gender (Riding, 1997)). This caution, as regards the feedback participants is tempered by the correlation analysis of the ratings patterns in the pilot and the main study participant groups (10.5.2); i.e. there are two studies and their results on the feedback side support each other. On the designer side, the designer participant group would probably not be considered representative of all professional designers as they were student interior designers.

However, the participants' experience of this new form of visual communication does provide a window into the likely appeal of the visual feedback formats for both feedback users and designer users.

10.5.1 The Feedback Givers

Results from the main study feedback task VAS item scores showed the following:

In line with the visual-verbal cognitive style dimension there is evidence of two groups within the feedback participants: image-likers and text-likers, as defined by their stated preference of format and confirmed by the pattern of different mean scores for the two groups on the two VAS items over the three answer formats (Figure 10.7).

Additionally, the results show that, for Utility:

1. Text-likers reported that they were better able to express their emotions using text and did not report images to be more fun to use than text.
2. Image-likers found the abstract images more useful for expressing their emotions than did text-likers.
3. Image-likers image likers reported no clear difference between their ability to express themselves with abstract images compared to text; we interpret this as image-likers thinking abstract images are as good as text for expressing their emotional reaction to designs.

Lastly, for Interest, the results show that:

4. Image-likers reported the image formats more fun to use than did the text-likers.
5. Also, although there was not a statistically significant difference between the means, the image-likers' mean scores for Interest indicates they found *abstract* images fun to use whereas the same cannot be said of text. (The text mean score shows they were equivocal as to whether text was fun or boring).

Qualitative themes from the post-task survey (Figure 10.8) showed the following:

6. Many of the participants (including text-likers) thought that *images* are good for expressing emotions. Conversely, a smaller number (half, which included image-likers) thought *text* is good for expressing emotions. A few thought both text and images are good for expressing emotions.

7. There was a substantial body of opinion (both image- and text-likers) that was dissatisfied with the emotion expressiveness of the Emotive204.
8. Some participants already think of themselves as visual people with three volunteering this self-categorisation unprompted.

Main Study Feedback Side Conclusion

Taken together, and notwithstanding the reservations about the makeup of the participant group, the above results are good evidence that a substantial proportion of people, *image-likers*, would a) enjoy using images chosen from perceptually organised browsers to express their emotional reactions to a design and b) consider those image selections as being as effective as text at expressing their emotions.

10.5.2 Pilot and Main Study Feedback Task Results Correlation

In Figure 10.10 the pilot data (normalised but not log transformed) are shown with the main study image-liker and text-liker data. The significance of this is discussed below.

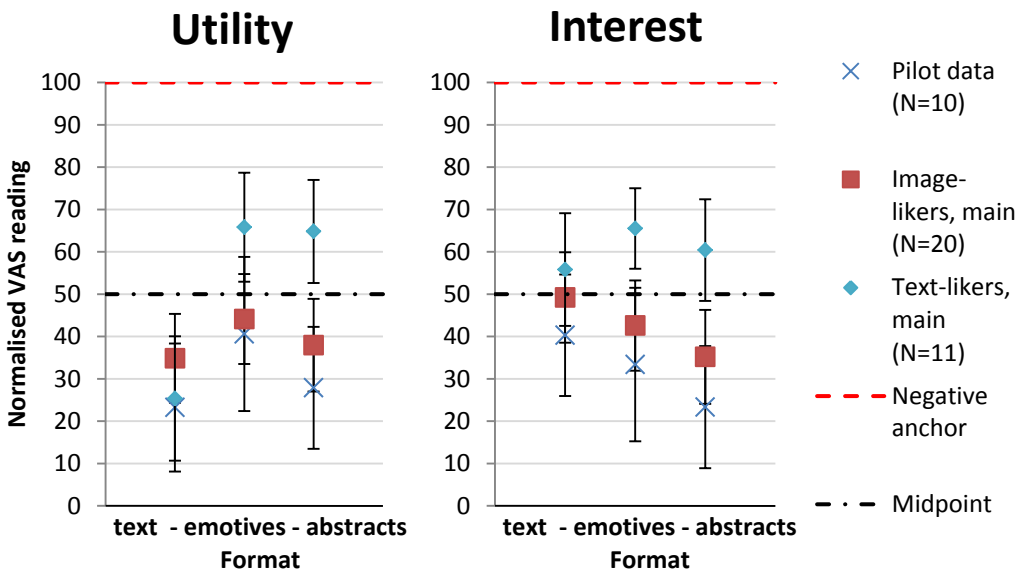


Figure 10.10 - Charts showing the correlation between pilot data and image-likers from the main study. VAS readings have been normalised 0-100 to allow comparison. The pilot data is shown here without log transformation. 0 marks the positive anchor. The correlations are Pilot data vs. Image-likers: strong (r = 0.95); twice as strong as Pilot data vs. Text-likers: medium (r = 0.47).

The VAS readings for Utility and Interest in the pilot were so skewed towards the positive end of the scales that they required to be log transformed to fit a normal distribution for parametric testing. However the main study VAS readings did not require this. The task conditions, participant group constitution, and VAS readings for the pilot and the main study are compared in Appendix C p.214. A correlation analysis (reported in the caption to Figure 10.10) showed that that the pilot participants behaved like the image-liker group from the main study. This may be more significant than just providing evidence that the VAS items for Utility and Interest are stable. The pilot group were all fashion students; i.e. creative people. If we theorize, equating the image-likers from the main study feedback participants with people of visual cognitive style, we can say that the pilot study participants are also likely to be visual individuals. It may be that creative people are over represented in the proportion of the population to whom the CVFM may appeal. If this is the case then the visual crowd feedback may have a higher value to designers than text-based feedback coming from the general population.

Main Study and Pilot Feedback Side Correlation Conclusion

a) The pilot participants behaved like the image-liker group from the main study and may have consisted almost entirely of such "image-likers".
b) The fact that the differences between the pilot and the main studies can be explained in this way provides evidence that the VAS items used to measure Utility and Interest in the two studies are stable. It also provides further evidence that, as far as feedback users of the method are concerned it may just appeal to one portion of the general population (image-likers).
c) The positive correlation between creative people and engagement with the visual feedback format as crowd users has implications for the quality of feedback for designers from the CVFM.

10.5.3 Designer Participants Receiving the Feedback

Even with the caveat of the participant group being student interior designers and not necessarily representative of professional designers, the fact that eleven of the twelve participants would be enthusiastic users of a service offering the CVFM shows that designers value this new form of visual feedback. It is also likely there is an element of

hunger for any form of feedback (not just visual) in their wish to consume the new formats.

However, from the themes there is clearly much in the visual feedback formats that ***cannot*** be offered by conventional text feedback. In particular the reverse mood board function of the abstract image format; but also: the non-threatening nature of the abstract feedback summaries while still provoking a designer to rethink textures and colours in their design; the fact that, when it was possible to compare the immediate inspiration from the separate feedback formats, 0/3 were inspired by text and 5/8 were inspired by image feedback; the ability of the emotive format to convey negative feedback in a way that some designers find less affecting than text; the ability to make feedback-givers focus on their emotional and perceptual reaction to a design rather than stray into the conventional critique that is encouraged by text; and the visual ideas to spur development in prototypes and to take forward to new projects.

The ambiguity of images, seen as a disadvantage by participants who preferred text, was seen by others as a benefit allowing alternative interpretations including allowing feedback that might be considered down-beat or negative to be overlooked in favour of more positive interpretations. (However, the negative feedback theme suggests that any negative image feedback would, in practice, be focussed on by designers).

From the themes and the format preferences it does appear likely that the visual-verbal cognitive style dimension is in evidence in the designer participants as it was in the feedback participants. This might be considered surprising as one might assume designers would be highly visual people. Perhaps here the nature of the participant group is a factor? They are student interior designers. It is not known yet if they will all go on to become professional designers.

Designer Participants Conclusion

a) Overwhelmingly, the designer participants desired to use a service offering the CVFM
b) For designers the CVFM offers several benefits not available from text feedback, including: the perceived mood via abstract image feedback and its non-threatening nature; helping the feedback crowd to focus on emotions instead of critique; freedom to avoid downbeat or negative feedback; and the visual inspiration not present in text.

10.5.4 The Two Image Types (Abstract and Emotive)

As to which type of images, abstract or emotive, would be preferred by the feedback participants in the future, a firm conclusion may not be possible. In this evaluation with the two alternative image sets as they were constituted, the Abstract500 appears to have been received more favourably than the Emotive204. On the feedback side there were generally poorer VAS item scores for emotives and themes expressing dissatisfaction with the choice of images in the Emotive204.

However, there is the confounding factor of the different numbers of images in the two sets, 500 in one and 204 in the other. The smaller number in the Emotive 204 was due to the pursuit of balanced numbers across the 19-term *design feedback emotion subset*. For some terms there were not enough good images (by emotion profile) in the full Emotive2000 to allow more than 10 per term. Notwithstanding this, discarding quantitative balance and allowing in further good images for terms where these existed in the full Emotive2000, may have been a better strategy as far as the experience of feedback and designer participants was concerned.

Alternative filtering could have been applied to provide for a graduated selection (by emotion intensity and mix) for some terms, thus addressing the concerns about lack of moderation in the emotions (e.g. some slightly less joyful "joy" images, perhaps with a joy *and* serenity in their profile.). Thus, allowing 500 emotion images would have removed the confounding factor of numerical difference and allowed a fair comparison between abstract and emotive imagery.

On the other side of the argument, however, this might have affected the amount of negative feedback perceived in the emotive image feedback by the designer participants. Too few negative images may have prevented some themes (about negative feedback being perceived in the emotive image summaries) being expressed by designers and prevented them coming to light in the evaluation. On the other hand, too many negative images might have skewed the balance of the feedback to become overly negative and put designers off the whole idea of image feedback. However, of the designer participants who preferred image feedback, most preferred the abstracts. This is despite experiencing them through the 10-image summaries equally for both abstracts and emotives; i.e. to the designers, in the feedback, the abstracts and emotives appeared equal in quantity.

Thus, it does appear that the mainly abstract imagery of the Abstract500 (which did not feature human faces) was generally more appealing than the Emotive204 for both expression and reception of emotion feedback.

10.5.5 The Evaluation Research Questions Revisited

In Table 10.6 the evaluation research questions posed at the start of Chapter 9 are revisited to establish whether they have been answered.

ERQ No	*Evaluation Research Questions* with Answers
	Feedback givers (the crowd)
1	*Do feedback givers prefer using images or text when describing their emotions?*
	Some do. Some don't. In the feedback participants there were two groups: text-likers (11) and image-likers (20)
2	*Do feedback givers find the image formats more engaging than text?*
	The image-likers found the abstracts fun while being equivocal about text. However the difference is not statistically significant. Text-likers showed no difference between the fun of text and images.
3	*Do feedback givers feel able to express their answer using the image formats?*
	Image-likers reported they were equally able to express their emotions with images (particularly abstracts) as with text. However, text-likers judged the Utility of text to be better than images.
4	*Do feedback givers feel more or less inhibited in expressing their emotions using images compared to text?*
	More participants reported holding back when using text (5) than with images (1).
	Designers (those consuming the feedback)
5	*Do designers value the image feedback formats?*
	Yes.
6	*Do designers prefer receiving feedback about emotions using the image formats or text as the medium?*
	Some participants preferred images (7). Some preferred text (5).
7	*Are designers inspired by the visual feedback to make changes to their designs?*
	Some were. A quantitative analysis after viewing first feedback showed inspiration from images but none from text. The development status of the design influenced this. Designers presenting prototypes were more likely to be inspired to change.
8	*What do designers think of the image formats as a method of feedback about emotions experienced by viewers of their designs?*
	Participants expressed the view that images helped the crowd focus on emotions rather than stray into standard critique.
9	*What do designers think of this method of communication?*
	Abstracts were thought useful, particularly at revealing the perceived mood of the design. Abstract images were more popular than emotives.
10	*Would designers use a service providing the visual feedback formats?*
	Yes. (Overwhelmingly so).

Table 10.6 - Evaluation research questions revisited.

10.5.6 The Possibilities for a New Visual Feedback Service

The use of a service providing the visual feedback formats was discussed with the designer participants and their reaction as a group (Theme 6) suggests it would be popular among designers. One was particularly animated about the prospect: "*I'd love that! I'd absolutely love that yeah!*" *[D8]*. They all agreed that they would get involved in feedback cycles by putting forward prototypes and responding to feedback with developments in their designs (rather than just putting up finished designs). One possibility for such a service would be in building a following for a designer by segmenting the crowd. When this scenario was put to one participant and she responded, *"That's a million dollar idea! You should get an app!"[D1]*.

Chapter 11

Summary and Conclusion

The last chapter, in Section 11.1, summarises what the thesis has achieved with reference to the goals stated at the outset. Section 11.2 summarises the results obtained while satisfying these goals. Section 11.3 discusses the implications of the findings, suggests possible directions for future work and hypothesises about how crowds could be engaged in visual feedback. Lastly, Section 11.4 will set out the thesis final conclusions.

11.1 How the Thesis Goals Were Achieved

This thesis began by proposing a novel method of obtaining crowdsourced intuitive visual design feedback (the CVFM) and setting the goals a) to develop the means to implement the method sufficiently to allow its evaluation, and then b) to evaluate it. It is now argued that these goals have been achieved. The rest of this section will summarise how this was done.

Chapter 2 put the case for there being a gap in the current provision of feedback modes and that this could be met by the CVFM using images as its medium. In particular, it was noted that a) some people think more visually than verbally and b) intuition is important in decision making (e.g. purchasing decisions) and emotions play a role. Chapter 3 selected a method for creating an intuitive browsing environment based on perceptual data to be used for the CVFM.

In Chapter 4, the Abstract500 image set consisting of 500 Creative Commons licensed abstract images was gathered and perceptual data was collected. The resulting similarity matrix was used to inform a self-organising map (SOM) browser presentation of the image set ready for use by feedback participants in evaluation studies (Figure 4.5).

In Chapter 5 it was established that the CVFM required that image summarisation be applied to a crowd's image selections to create concise reports for designers. Existing work on image summarisation was examined and although one candidate method was

identified, it was not ideal. It was decided to specifically develop a summarisation method which exploited the Abstract500's perceptual similarity matrix. Chapter 6 described the development of an image summarisation algorithm and a prototype implementation was created for use in the evaluation studies.

Chapter 7 described a two-stage experiment in which participants chose images from the Abstract500 browser to represent terms and then another group of participants rated the images (and summaries made from them) for each of the meanings, allowing communicative effectiveness to be assessed (Figure 7.7). The Abstract500 was found to be better for communicating descriptive terms rather than emotive terms. It was also shown that, on the whole, the visual summarisation (using 10 representative images) was effective at preserving the intended meaning of image selections.

Thus Chapter 7 had shown that the visual summarisation worked with the Abstract500 browser. However, as Chapter 7 had also shown that the Abstract500 browser was less effective for emotion terms than for descriptive terms it was decided to develop a further image set more suited to emotions. In Chapter 8 a model of emotion was selected and a subset of 19 of its terms identified as being suitable for design communication. 2000 Creative Commons images associated with the 19 terms were gathered and then shown to paid crowdsourced participants who tagged them with terms from the model. This produced a normalized emotion tag frequency profile representing the judgments of 20 participants for each of the 2000 images (Figure 8.3). Using these profiles, the set was filtered to the Emotive204 (Figure 8.9), 204 images (emotives) balanced over the subset of 19 terms. The emotives were arranged in a SOM browser defined by the emotion tag frequency profiles in a similar way to the Abstract500 browser (based on perceptual similarity data).

Thus, Chapters 2 to 8 produced the major components enabling evaluation of the CVFM: 1) two image browsers (abstract and emotive) specially constructed to allow intuitive image selection by crowd users and 2) image summarisation to condense high volumes of image selections from a crowd for presentation to designer users. By-products of this were two data sets: the Abstract500 image set with associated 500x500 perceptual similarity matrix, and the Emotive2000 image set with associated emotion profiles (normalised emotion tag frequency vectors).

Chapters 9 and 10 described two evaluations of the CVFM, a pilot and the main evaluation. The evaluations were considered to have two sides 1) the feedback side and

2) the designer side. On the feedback side feedback participants representing the crowd viewed designs, were asked “How did the design make you feel?”; then, to answer, they chose two types of images (abstracts and emotives) and entered text. While doing this they rated the three formats (abstracts, emotives and text) for both utility and interest. On the designer side designer participants put forward designs, and then contrasted abstract and emotive image feedback summaries and text during semi-structured interviews revealing what they thought of them.

Thus Chapters 2 to 8 developed the means and Chapters 9 and 10 did the evaluations of the CVFM, satisfying the goals of the thesis. The next section summarises the results from the evaluations.

11.2 Summary of Results

The feedback participants showed that they had behaved as two groups, image-likers and text-likers, through a) their ratings of abstracts, emotives and text for *utility* and *interest* and b) their answers in a post-task survey.

A correlation analysis showed that the pilot feedback participant group, all creative people, behaved as image-likers.

When rating *utility*, text-likers reported that they can express themselves better with text than images; whereas image-likers reported that abstract images are more useful for expressing their emotions than reported by text-likers. Indeed image-likers reported no difference between the usefulness of abstract images and of text for expressing their emotional reaction to designs.

When rating *interest*, image-likers found the image formats more fun to use than did the text-likers. Indeed image-likers reported that the abstract images were fun whereas they rated text as neither fun nor boring.

These feedback participant results suggest that there is a section of the population who would find the CVFM fun to use and who think that images are as good as text for communicating their emotional reaction to a design. Additionally the results suggest that creative people may be more likely to be image-likers and engage with the CVFM, raising the prospect of gaining high quality inspiring feedback for designers which, without the CVFM, would not be collected.

11 out of 12 designer participants said they would be enthusiastic users of a service which allowed them to upload their designs and receive feedback in the new visual formats. They were inspired by image feedback with changes being motivated where, comparably, text feedback motivated none. They were aware of ambiguity in images but freely interpreted the image feedback assigning messages to images and groups of images on feedback summaries. Participants reported that the abstract image summaries could act as "reverse-engineered" mood boards reflecting the crowd's perception of the mood of a design. Also, although able to read negative feedback in both *text* and in *emotive* images, designer participants found the *abstract* images could be inspirational without being perceived as threatening or negative. In addition, they reported that, in their view, images had made the feedback participants focus better on emotions instead of drifting into a conventional critique, neglecting emotions, as they did with text.

Taking the feedback and designer side results together, we have evidence that the designer participants and image-likers among the feedback participants think they can communicate using the CVFM. The correlation of the pilot study group with the image-likers from the main study feedback participants, and the fact that cognitive styles have been shown to be independent of age, gender and intelligence (Riding, 1997) mean that it is reasonable to suggest that the findings apply beyond the participants in the two studies.

11.3 Implications and Future Work

In this section, questions raised by this thesis, but outside its scope, are discussed as possible areas of future investigation. The section ends by speculating about how crowds might be engaged in visual feedback.

11.3.1 The Imagery and Summarisation

Perceptually organised abstract imagery, such as the Abstract500, can be used as a medium to access the perceived mood of a design and portray it in a form already accessible by designers without any prior new acclimatisation or familiarisation with the format. The presentation and summarisation of the Abstract500 can be deemed to have worked successfully to enable communication and was embraced by both visual crowd members and designers.

Perceptually organised emotive imagery as a feedback medium has the potential to help feedback givers focus on their emotions. The nuances of the emotive image feedback would require further development for the majority of designers to find it acceptable. These nuance factors include: the balance of emotions depicted within the imagery; also interactive access to cluster component images in summaries could usefully be added. In Chapter 7, this thesis investigated the communicative effectiveness of a) the Abstract500 browser and b) the summarisation algorithm finding that the Abstract500 browser had varying communicative effectiveness and did not perform well for emotion terms but the summarisation algorithm, on the whole, was effective. A similar experiment could be done to a) confirm that a browser populated with images from the Emotive2000 performs well at communicating emotion terms and b) that the summarisation algorithm is effective when applied to an image set which has emotion tagging frequency vectors instead of similarity data associated with it.

11.3.2 Cultural Considerations

While some aspects of emotion in imagery are considered universal and thus bridging cultures, such as some facial expressions (Plutchik, 2003) (Ekman, 1984) (Darwin 1965), other aspects of imagery, such as colour, can vary between cultures in their emotional associations (McCandless, 2009). Also, there are subtle cultural differences in interpretation of the "universal" facial expressions (Yuki et al., 2006). Images provide for non-verbal communication which should be language independent and thus have an advantage over text but intercultural differences may need to be taken into account. An investigation of cultural differences in the interpretation of the image banks built for the evaluation would help improve the formulation of further image banks.

In addition, as was noted in 2.6.3, there is evidence that cognitive styles may vary with culture (at least in the case of the holistic-analytic dimension). Thus it might be that the appeal of the CVFM may also vary with culture and this too could be investigated.

11.3.3 Cognitive Styles, Intuition and Emotion

We can theorise that the image-likers and text-likers in the main study participant group equate to individuals who are either more visual or more verbal in cognitive style. The measurement of cognitive styles in participants was out of scope for this thesis. However, it might be possible to prove or disprove this theory by repeating the main

evaluation including an additional form of measurement to assess cognitive style within participants.

The result showing that the pilot feedback participant group (all fashion students) correlated with the image-likers from the main study and are therefore people who are more likely to be engaged by the CVFM raised the prospect that the CVFM may disproportionately attract creative users into giving feedback. Confirming this and investigating differences in feedback from participants who are more creative and those who are typical of the general population would be interesting.

If the CVFM does encourage use of intuition and emotion, this might have implications for the quality of feedback obtained in general and from different age groups. As mentioned already in 2.6.4, for complex decisions, "going with one's gut" and not over-thinking a decision has been shown to produce superior outcomes compared to a deliberative approach (Mikels et al., 2011). With regard to the age of potential crowd users (feedback givers), the feedback participants in the two studies in this thesis were relatively young people. However, Mikels et al. (2010) showed that older people make better quality decisions when using feeling focused decision strategies compared to detail focused strategies. This raises the possibility that intuitive and emotion based visual feedback from older people (encouraged by the CVFM) might be of superior quality compared to feedback they may otherwise contribute using traditional text methods which encourage deliberative thought. With the aging of the population now a common fact in industrialised counties (Fendrich & Hoffmann, 2007) this will be a growing consideration.

11.3.4 A New Service and Crowd Engagement

The finding in Theme 6 (Table 10.3) that, overwhelmingly, the designer participants wished to use an internet service offering these visual formats, shows a potential market for such a tool among designers. Sub-theme 6.2 showed that the designer participants were unanimous that the best use of such a service would be in developing and refining a prototype design via cycles of crowd feedback. Such visual feedback cycles if recorded could constitute an attractive design narrative adding value to any final product (2.3.4). In addition, if the crowd involved in feedback can be cultivated as a virtual customer community (2.3.3) allowing a designer to build a following this would bring financial gain to join any creative gain from the CVFM for designers; this possibility

was clearly appreciated by the designer participants (10.5.6). Thus the designer participants saw dual potential in the CVFM.

How a crowd might be engaged in giving feedback using the CVFM is an open question. Social networks can be a useful source of feedback on ideas (Dow et al 2013) and could be a good conduit through which designers could use the new mode to leverage participation. An idea or piece of news can spread rapidly through social media given the right circumstances (Kaplan & Haenlein, 2001) or, like the uptake of some social network based applications, can spread only in a limited way and then reach a plateau (Kirman et al., 2010)(Nazir et al., 2008); this can occur if it fails to spread beyond a few peoples' immediate social network which are usually limited in size (Hill & Dunbar, 2002). Thus introducing it as a social network based application might be problematic.

However, as a new visual mode of communication it could perhaps be adopted to work alongside existing text feedback methods as an alternative for users of a more visual cognitive style. This visual alternative to the text input field could be offered within the domain of design feedback e.g. feedback communities such as Dribbble (2015) or it could be an entertaining alternative to the text field and "Like" buttons in services such as Facebook (2015), YouTube (2015) and Instagram (2015). Indeed photo sharing social media services are likely to be frequented by users already open to responding visually. Additionally, were the service to become popular, committed users might enjoy being involved in the development and expansion of the imagery by sourcing and categorising images, adding a further social aspect to belonging to the visual crowd.

11.4 Summary of Thesis

In this thesis a method of obtaining intuitive, perceptual, image based, design feedback from a crowd, and summarising it for consumption by designers was proposed (the CVFM). The two major components needed to evaluate the method, an intuitive abstract image browser and image summarisation based on perceptual similarity data were developed. The effectiveness of these for communication was assessed experimentally. A further image browser populated with emotive images based on crowdsourced tagging was developed to offer improved emotion communication.

These components were used to evaluate the CVFM. The visual feedback formats were well received by designer participants and they desired to use a service offering this new style of crowd communication. Feedback participants, representing the crowd, behaved as two groups, image-likers and text-likers. While the text-likers did not particularly value the CVFM, the image-likers thought using it was fun and an effective way to communicate the emotional impact of designs. Correlation between the rating behaviour of a group of pilot feedback participants (all creative people) with the image-liker group in the main study a) reinforces the results by confirming the stability of the methods used and b) raises the prospect of the CVFM appealing to creative people in particular.

The main achievement of this thesis is in showing that crowdsourced intuitive visual design feedback, a new form of social computing interaction based on images, can work for designers and would be engaging for a section of the population to take part in giving visual feedback.

Appendix A Development of an Algorithm for Visual Summarisation

This appendix accompanies Chapter 6. It begins with a brief reference to the summarisation algorithm code. The remainder of Appendix A is taken up with the *Dimensionality Reduction Exploration* which describes that exploration under headings: *Aim*, *Method*, *Images and Categories of Interest*, *Results*, and *Conclusion*.

Code for the Algorithm

An implementation of the algorithm can be found via the appendix for Chapter 10.

Dimensionality Reduction Exploration

To inform the choice of dimensionality reduction method, an exploration of the Abstract500 image set and its similarity matrix was carried out as reported below.

Aim

To explore dimensionality reduction for visualisation and summary construction.

Method

1) MATLAB scripts were created to produce 3D visualisations of the similarity matrix, which are interactive in a web browser with appropriate *X3D* plugin. Particular images and categories of images were identified. How these images and image categories were positioned within the visualisations was compared. Screenshots illustrating the position within the view of the images of interest were taken. The situation of the images of interest within the visualisations was noted.

   Four dimensionality reduction methods were applied: a) Classical MDS b) Non-metric MDS c) Isomap and d) IsomapII. (See Chapter 6 for references). For IsomapII the 100 bootstrap images were used as the "Landmark" data points.

2) The amount of variability in the data encompassed by the three visualised dimensions was assessed through the stress and residual variance plots generated during the production of the visualisations when the given reduction methods were applied to the data.

Images and Categories of Interest

1. Man-made/structural; (See Table A1 for example images)
2. Nature/Botanical (See Table A1 for example images)
3. Abstract multi-coloured patterns
4. Grainy/Gravely/Rocky/Rusty
5. Diffuse colourful
6. Singleton in bootstrap sort (image #10). This image stood out as often a singleton during the bootstrap card sort. (See Table A1)

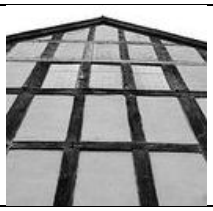
Man-made/structural				
Image Expt. No	25	495	118	204
Image				
Nature/botanical				
Image Expt. No	64	318	110	28
Image				
Singleton in bootstrap sort				
Image Expt. No	10			
Image				

Table A.1 – Examples of images in the categories of interest

Results

Firstly the visualisations:

All of the categories 1 to 5 appeared as clusters which appeared coherently within a defined region of all of the visualisations. See Table A.2. The non-metric MDS appeared the more open view.

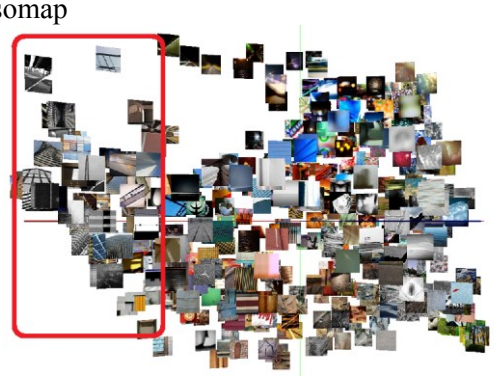
Classical MDS 	Non-metric MDS 
Isomap 	IsomapII 100 Landmarks 

Table A.2 - The location of the "Man-made/structural" region in the 3D visualisations.

Image 10, often a singleton in the bootstrap card sort for the Abstract500, was best represented as an outlier by the non-metric MDS visualisation. See Table A.3.

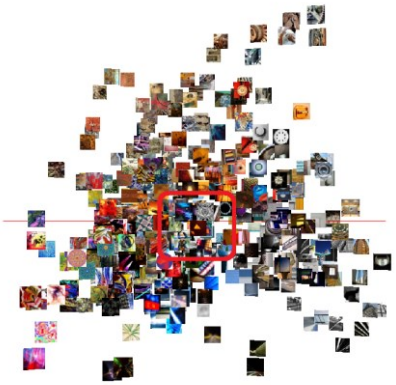
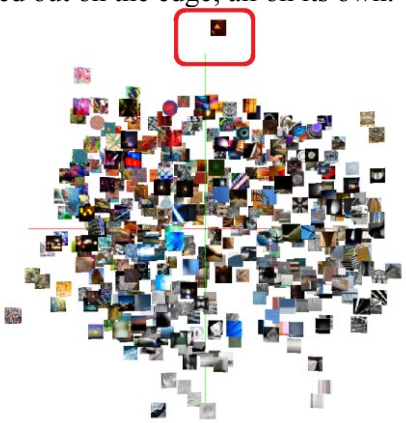
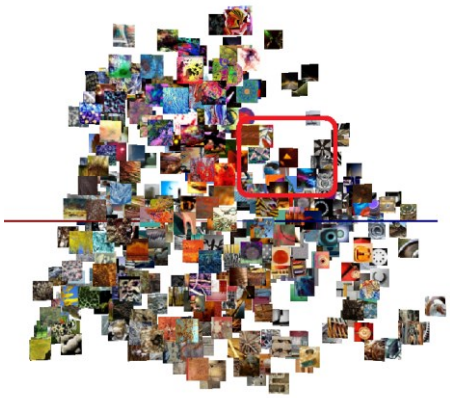
Classical MDS – Not on the surface, but in amongst other images. 	Non-metric MDS – Isolated out on the edge; all on its own. 
Isomap On the surface but within a pocket/hollow in the distribution. 	IsomapII 100 Landmarks Like Isomap; on the surface but within a pocket/hollow in the distribution. 

Table A.3 - The location of the singleton image 10 in the 3D visualisations

Secondly, the charts describing the variability encompassed by the three dimensions portrayed by the visualisations:

Figures A.1, A.2, A.3 and also Figure 4.3 (p. 52) (for classical MDS) show that all four reduction methods result in the first 3 dimensions describing a large proportion of the variability in the data.

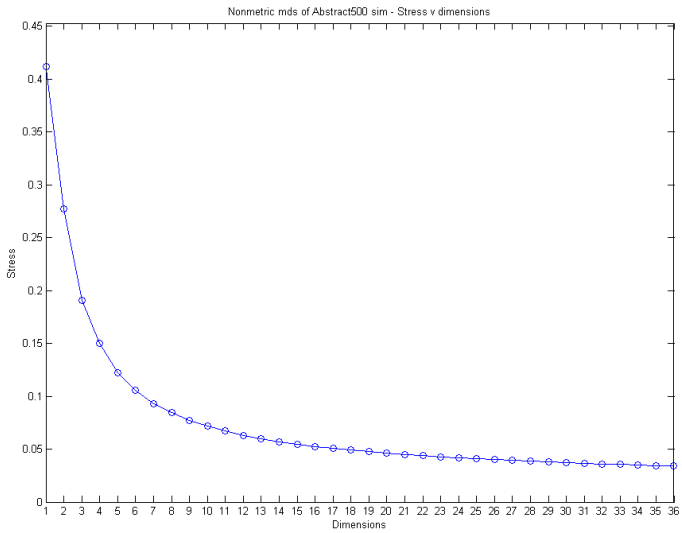


Figure A.1 - Plot of stress vs. dimensions for non-metric MDS of Abstract500sim.

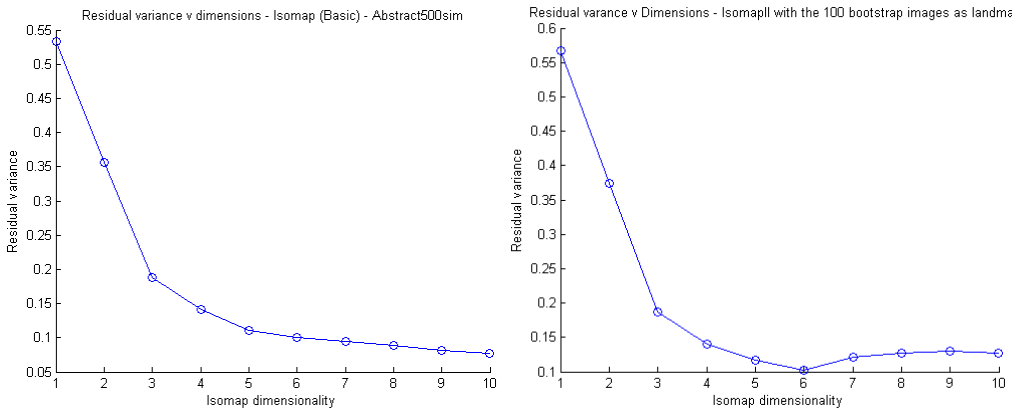


Figure A.2 - Plot of residual variance v dimension for Isomap reduction (left) and Isomap (Landmark) (right)

Conclusion

The distributions in the different views are not vastly different. The non-metric MDS provides the more open view. That view is the one that gives a better representation of the isolation of image 10.

The residual variance and stress plots tend to confirm that, in general, three dimensions from each reduction method successfully describe much of the variability in the perceptual similarity data of the *Abstract500*.

Due to the better portrayal of the singleton image 10 as an outlier and the more open structure of the non-metric MDS view this is the one chosen as the reduction method for the summarisation.

Appendix B Evaluation Study Pilot

This Appendix accompanies Chapter 9.

Administrator's Intro Script

Today I will be showing you some pictures of designs and asking you to comment on them responding to a specific question using different answer formats.

The first design you will see and respond to is for practice so that you get to see and use all 3 different answer formats. You also get to experience the questions which seek your judgements about the 3 different answer formats.

Next you will be shown a number of designs and after viewing each one you will be asked to respond using one of the answer formats.

Then you will be asked to give your judgements about that answer format.

You then view and comment on the designs using the other answer formats and give your judgements about them in a similar way.

All your responses will be stored and processed anonymously.

Your comments on the designs will be collated along with other participants' comments and summarised. The designers whose designs you commented on will be shown the summaries of the collated feedback. (But the designers will not know who gave the comments).

[The appropriate consent form was then completed and signed by the participant]

Administrator's Debrief Script

Are you finished?

Now you are finished I need to tell you that your comments may not actually be summarised and shown to the designers. In this pilot study it is unlikely that this will be possible. It was necessary that you thought the designers would definitely see a summary so that we could find out what you thought about your freedom of comment using the different formats.

Now you are finally finished is there any comment you have?

Thank you!

Feedback Participant Task/Questionnaire Example (here reduced to A5 from A4)

NB: Abstract images, emotive images, and text were labelled L, P, and Q respectively. The format presentation order (in this case Q-P-L) was randomised. See *Record of the Randomised Format Order* (which follows the questionnaire in this appendix).

Training Phase

1. View the design on Computer screen
2. Using the computer and **Format-Q** please answer this question:
   "How did the design make you feel?"
3. Using the computer and **Format-P** please answer this question:
   "How did the design make you feel?"
4. Using the computer and **Format-L** please answer this question:
   "How did the design make you feel?"

Please answer the following questions by marking an X on the line to indicate your opinion. You may place your mark at either end, or anywhere in between. For exan you were asked "How warm do you feel?", you might answer like this:

Hot _____ **Cold**

5. About Format Q:
   When using Format Q:
   a) How well were you able to express yourself?

Completely _____ **Not at all**

   b) In relation to freedom of expression i.e. freedom to say whatever you wanted without caring what anyone, including the designer, might think about the answer you gave: How free did you feel in giving your answer?

Totally _____ **Totally**
Free **Inhibited**

   c) How interesting was this way of giving your answer?

Very Much _____ **Very much**
Fun **Boring**

6. About Format P:
   When using Format P:
   a) How well were you able to express yourself?

Completely _____ **Not at all** ☐

   b) In relation to freedom of expression i.e. freedom to say whatever you wanted without caring what anyone, including the designer, might think about the answer you gave: How free did you feel in giving your answer?

Totally _____ **Totally** ☐
Free **Inhibited**

   c) How interesting was this way of giving your answer?

☐
Very Much _____ **Very much**
Fun **Boring**

Feedback Participant Task/Questionnaire (continued)

7. About Format L:

When using Format L:

a) How well were you able to express yourself?

Completely _____ **Not at all** ☐

b) In relation to freedom of expression i.e. freedom to say whatever you wanted without caring what anyone, including the designer, might think about the answer you gave. How free did you feel in giving your answer?

☐

Totally Free _____ **Totally Inhibited**

c) How interesting was this way of giving your answer?

☐

Very Much Fun _____ **Very much Boring**

Using Format-Q

Design 1 (out of 3)

1. View the design on Computer screen ☐
2. Using the computer and **Format-Q** please answer this question: ☐
   "How did the design make you feel?"

Design 2 (out of 3)

3. View the design on Computer screen ☐
4. Using the computer and **Format-Q** please answer this question: ☐
   "How did the design make you feel?"

Design 3 (out of 3)

5. View the design on Computer screen ☐
6. Using the computer and **Format-Q** please answer this question: ☐
   "How did the design make you feel?"

Please answer the following questions by marking an X on the line to indicate your opinion.

7. About Format Q:
   When using Format Q:

a) How well were you able to express yourself?

Completely _____ **Not at all** ☐

b) In relation to freedom of expression i.e. freedom to say whatever you wanted without caring what anyone, including the designers, might think about the answer you gave: How free did you feel in giving your answers?

Totally Free _____ **Totally Inhibited** ☐

c) How interesting was this way of giving your answers? ☐

Very Much Fun _____ **Very much Boring**

Using Format-P

Design 1 (out of 3)

1. View the design on Computer screen ☐
2. Using the computer and **Format-P** please answer this question: ☐
   "How did the design make you feel?"

Design 2 (out of 3)

3. View the design on Computer screen ☐
4. Using the computer and **Format-P** please answer this question: ☐
   "How did the design make you feel?"

Design 3 (out of 3)

5. View the design on Computer screen ☐
6. Using the computer and **Format-P** please answer this question: ☐
   "How did the design make you feel?"

Please answer the following questions by marking an X on the line to indicate your opinion.

7. About Format P:
   When using Format P:

a) How well were you able to express yourself?

Completely _____ **Not at all** ☐

b) In relation to freedom of expression i.e. freedom to say whatever you wanted without caring what anyone, including the designers, might think about the answers you gave: How free did you feel in giving your answers?

Totally Free _____ **Totally Inhibited** ☐

c) How interesting was this way of giving your answers? ☐

Very Much Fun _____ **Very much Boring**

Using Format-L

Design 1 (out of 3)

1. View the design on Computer screen ☐
2. Using the computer and **Format-L** please answer this question: ☐
   "How did the design make you feel?"

Design 2 (out of 3)

3. View the design on Computer screen ☐
4. Using the computer and **Format-L** please answer this question: ☐
   "How did the design make you feel?"

Design 3 (out of 3)

5. View the design on Computer screen ☐
6. Using the computer and **Format-L** please answer this question: ☐
   "How did the design make you feel?"

Please answer the following questions by marking an X on the line to indicate your opinion.

7. About Format L:
   When using Format L:

a) How well were you able to express yourself?

Completely _____ **Not at all** ☐

b) In relation to freedom of expression i.e. freedom to say whatever you wanted without caring what anyone, including the designers, might think about the answers you gave: How free did you feel in giving your answers?

Totally Free _____ **Totally Inhibited** ☐

c) How interesting was this way of giving your answers?

☐

Very Much Fun _____ **Very much Boring**

And Finally

1.
   a. Did you understand all the questions in this questionnaire? Yes / No
   b. If not please describe your difficulty here or verbally tell the experiment administrator.

2. If you have any comment or anything to add please use the space below

Thank you for your help in our research!

The Head Crowd Project Team

Record of the Randomised Format Order

Participant	Format order	Participant	Format order
1	L-Q-P	6	L-Q-P
2	P-Q-L	7	P-Q-L
3	L-P-Q	8	Q-L-P
4	P-Q-L	9	P-L-Q
5	Q-L-P	10	P-Q-L

Table B.4 - Pilot evaluation: record of format presentation order. Formats L, P, and Q were abstract images, emotive images, and text respectively.

Screens from the Interface

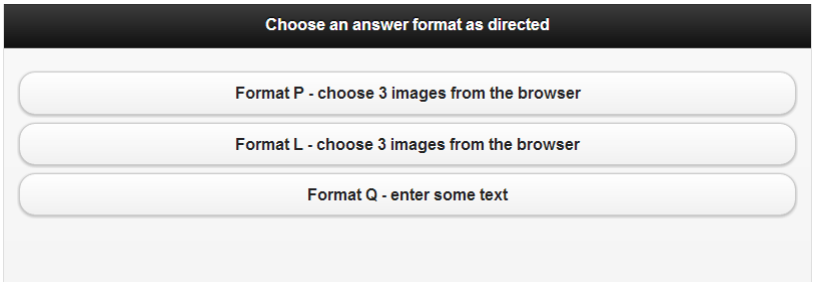


Figure B.1 - Interface main screen.

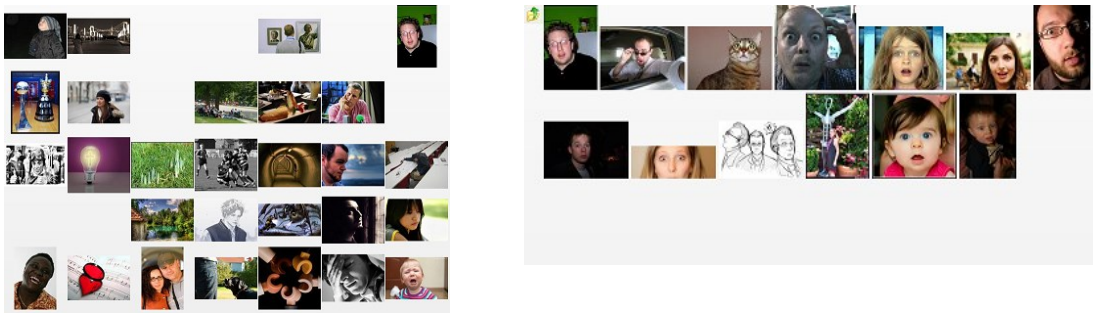


Figure B.2 - Emotive image format browser (Format-P). Top level (left) and an open stack (right).

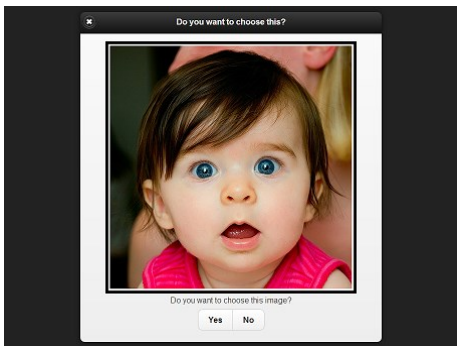


Figure B.3 - Confirm image choice dialogue (in this case for the emotive image format. There was a similar dialogue for the abstract image format).

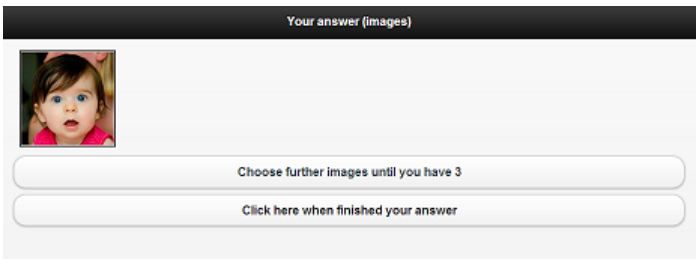


Figure B.4 - Choose further images dialogue for Format P. There was a similar dialogue for the abstract image format.

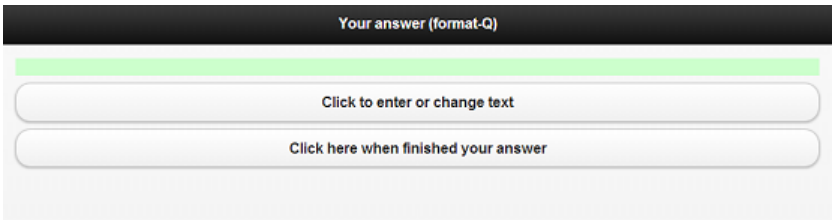


Figure B.5 - Text format first dialogue.

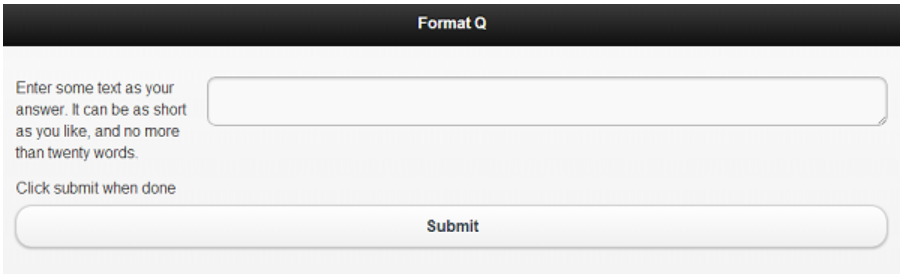


Figure B.6 - Text format second dialogue.

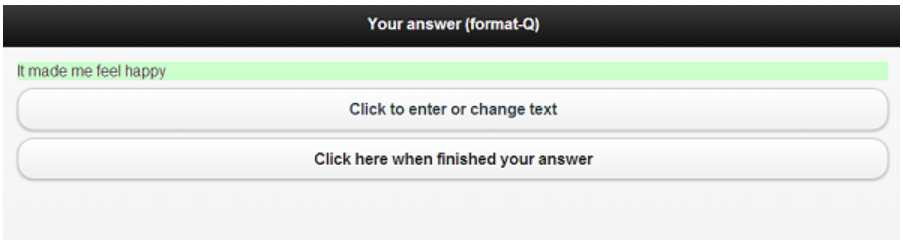


Figure B.7 - Dialogue after text entry .

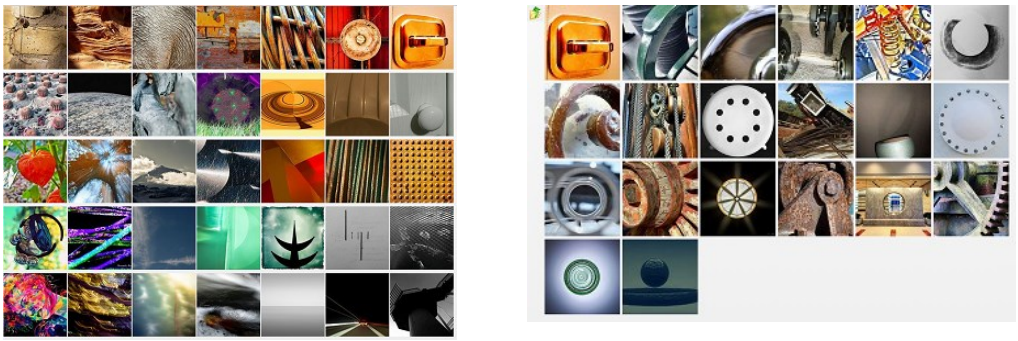


Figure B.8 - Abstract image format browser (Format-L). Top level (left) and an open stack (right).

Data Recording, and Validation

The following precautions were taken to aid accuracy in measurement of the participants' placement of their marks on the VAS item scales:

a) The length of the scale printed on all the questionnaires was 66.5mm (rechecked several times across randomly chosen pages and questionnaires).
b) The same ruler was used for all measurements.
c) Care was taken to start it at the same point and check the end was at 66.5 mm.
d) Measurement was to the nearest 0.5 mm as read from a constant view above the instrument.
e) The measurement was written onto the questionnaire next to each response.

The readings were entered into a spread sheet. The following list is a description of the consideration given to the scope for error and the table below details the steps taken to mitigate this.

The following were identified as possible sources of error:

1. Fine measurement: Millimetre level measurement: i.e. incorrectly measuring the response, not reading off at the correct millimetre or miss-positioning the ruler.
2. Gross measurement: Centimetre (actually half-centimetre) level measurement: misreading the gross part of the ruler scale when noting the measurement. This had been noticed while recording on a couple of occasions.
3. Number level transcription error i.e. miss-typing the number into the spread sheet.
4. Column entry transcription error. The format pages (P,Q,L) appeared in different orders on the questionnaires. So care had to be taken to enter the numbers in the correct spread sheet column group.

Source of Error	Mitigation during recording	Post data entry checks.
Fine measurement	Care taken, especially in placing the ruler. See above a), b) and c) above.	No further check
Gross measurement	Care taken after an error was noticed.	Manually, passed through the questionnaires doing a gross measure with the ruler and checking against the written value. 3 errors were found to be out by 5 mm. Only 1 error was in the experiment phase data. Corrected.
Number level transcription	Care taken	Manually, passed through the questionnaires individually checking the match between the sheet

error		and the questionnaire for number entry. (None of this error type encountered).
Column entry transcription error	Care taken	Manually, passed again through the questionnaires checking that the correct column sets were used for each format. One instance of this was found where Format L figures were interchanged with Format Q. Corrected.

Table B.1 - Steps taken to mitigate data recording errors in the evaluation pilot study.

All corrections were added to the data spread sheet allowing analysis to proceed.

Detailed Raw Results

The possible measurements ranged from 0 mm (positive) to 66.5 mm (negative). E.g. two of the most extreme values were: in Utility, participant 9 placed her mark on 0 for text format, indicating that she felt able to express herself "*Completely*" using text and participant 7 placed her mark on 50.5 mm along the scale for Emotive images format, indicating well along the scale towards "*Not at all*" and thus that she felt unable to express herself well using the emotive images (see table below for Item: Utility).

Participant	Text (Q)	Emotives (P)	Abstracts (L)
1	5	3.5	4
2	6	15.5	13.5
3	1.5	19.5	29
4	47.5	12	41.5
5	7	59	43
6	15.5	25.5	16.5
7	10	50.5	2.5
8	39	28.5	24.5
9	0	7	1
10	23	49	10

*Table B.2 - Pilot item: Utility: "How well were you able to express yourself?; Completely - Not at all". **Completely = 0** and **66.5 = Not at all**."*

Participant	Text (Q)	Emotives (P)	Abstracts (L)
1	8.5	4	3
2	8.5	19.5	20.5
3	21.5	6	14.5
4	23.5	9.5	15.5
5	5.5	8.5	59.5
6	8	29.5	15.5
7	6	51.5	1.5
8	19	46	26.5
9	1	29	0
10	3	1.5	2

*Table B.3 - Pilot item: Freedom: "How free did you feel in giving your answers?; Totally Free - Totally Inhibited. **Totally Free = 0** and **66.5 = Totally Inhibited**"*

Participant	Text (Q)	Emotives (P)	Abstracts (L)
1	6	6.5	5.5
2	9	17	21.5
3	59	4	15.5
4	31.5	10.5	28.5
5	33.5	62	52
6	32.5	28.5	9.5
7	14.5	48.5	1
8	25	18.5	8.5
9	22	4	1
10	35	22.5	12

Table B.4 - Pilot item: Interest: "How interesting was this way of giving your answers?; Very Much Fun - Very Much Boring. ***Very Much Fun = 0*** *and* ***66.5 = Very Much Boring****".*

K-S Tests for Normality on the Log Transformed Data

This was done using SPSS. See table below. For all 9 results distributions the significance value (Sig.) is not less than 0.05 indicating that none of them deviate significantly from normality (Field, 2009 p.246).

Distribution	K-S test statistic	df	Sig.(*p*)	Passes test?
Utility				
Text	0.15	10	0.20*	Yes
Emotives	0.14	10	0.20*	Yes
Abstracts	0.14	10	0.20*	Yes
Freedom				
Text	0.17	10	0.20*	Yes
Emotives	0.16	10	0.20*	Yes
Abstracts	0.26	10	0.06	Yes
Interest				
Text	0.19	10	0.20*	Yes
Emotives	0.14	10	0.20*	Yes
Abstracts	0.16	10	0.20*	Yes

Table B.5- K-S tests for the 9 pilot results distributions0.20 indicates that 0.20 is the lower bound of the true significance.*

The KS tests were done through the SPSS explore menu; the Lillifors Significance correction was applied (Field, 2009 p147).

Designer Interview - Order of Format Presentation

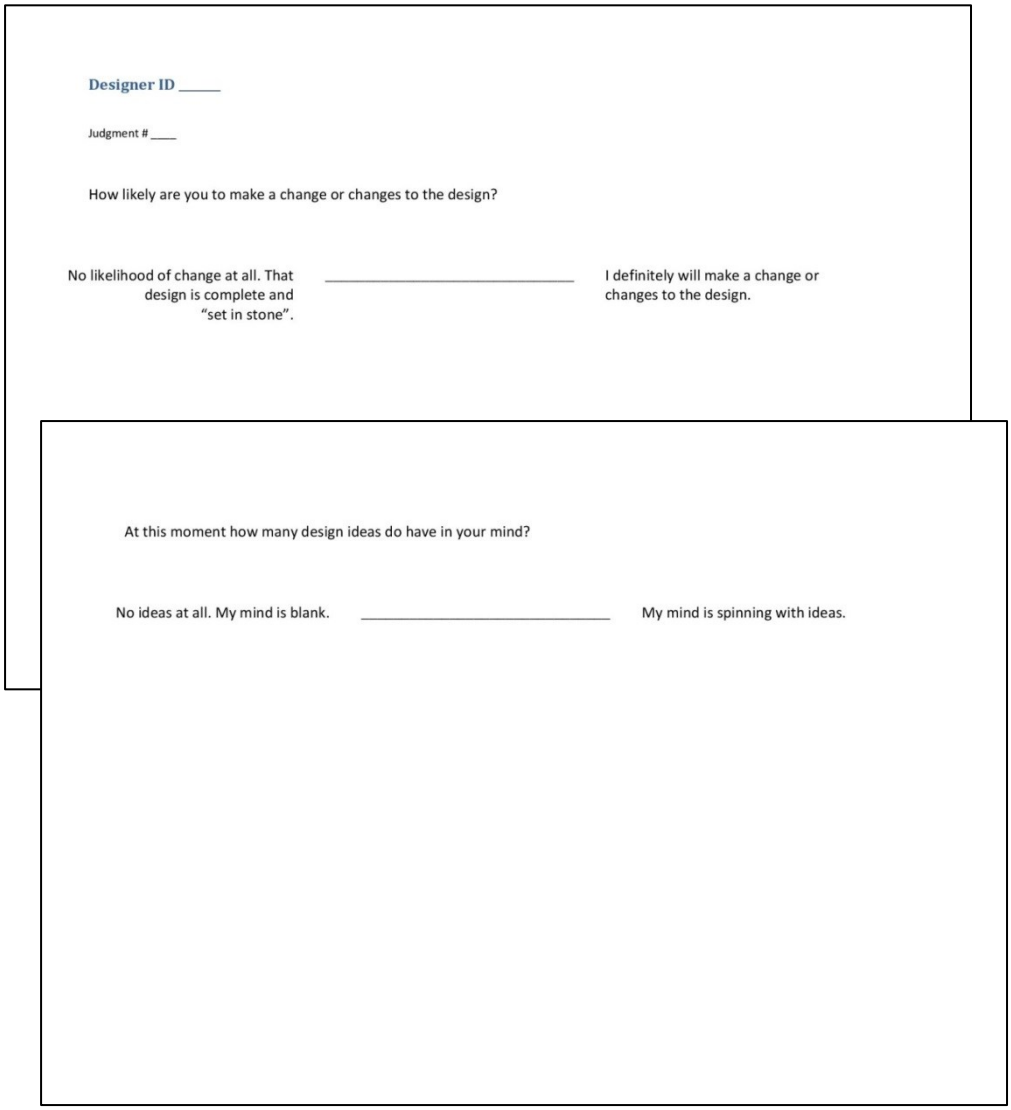
The order in which the feedback formats were presented was decided randomly. The random sequence generated was 1) Abstract images 2) Emotion images 3) Text.

Interview script

Intro	
Start clock	
Permission to record	
Purpose of research	
Format of interview	
As a designer, how do you use images in your work?	
Show image sets [laptop + iPad] - <u>free to interrupt me at any time</u>	
Allow explore on iPad and ask for speak aloud	
Confirm design image + Tell me something about your design….	
• Judgement 1	
<u>First</u> Reaction summary [iPad] and ask for think aloud.	
What do you think the participants are saying about the design? Is there anything unexpected? Anything that fits in with your design?	
What do you think of this format as a way to find out how people felt when looking at your design?	
One word to describe the summary?	
• Judgement 2	
<u>Second</u> Reaction summary [iPad] and ask for think aloud.	
What do you think the participants are saying about the design? Is there anything unexpected? Anything that fits in with your design?	
What do you think of this format as a way to find out how people felt when looking at your design?	
One word to describe the summary?	
<u>Third</u> reaction summary [iPad] and ask for speak aloud.	
What do you think the participants are saying about the design? Is there anything unexpected? Anything that fits in with your design?	
What do you think of this format as a way to find out how people felt when looking at your design?	
One word to describe the summary?	
General Qs	
Which reaction format did you prefer? - Can you say why?	
Which reaction format did you least prefer? - Can you say why?	
The question we asked the participants was "How did the design make you feel?" Is that the question you would have asked?	
Would you use an online service that allowed you to show your designs and get images as feedback? Can you fore-see any pros and cons of such a service?	

Designer Interview Judgement VAS Items

The two items were presented one after the other. This was done twice during the interview (See "Judgement" prompt on the interview script). They were A4 sheets.



Designer Interview Supporting Web Application Screens

Figure B.9 - Menu screen for designer interview supporting web application.

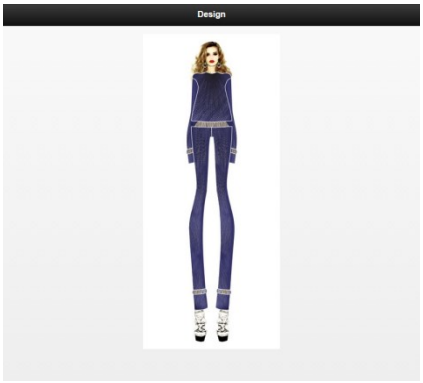


Figure B.10 - Design display pagefor designer interview supporting web application. Design image by permission DesPilot4

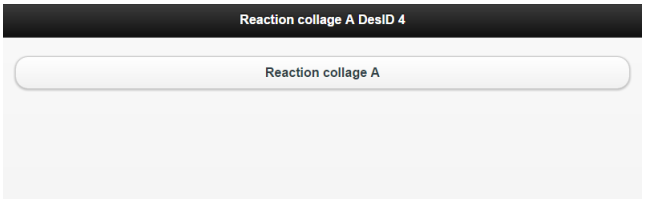


Figure B.11 - Example of intermediate screen; allowed the administrator to cue up a given feedback format for the designer participant to reveal on the iPad. This helped prevent the inadvertent revealing of stimuli out of sequence.

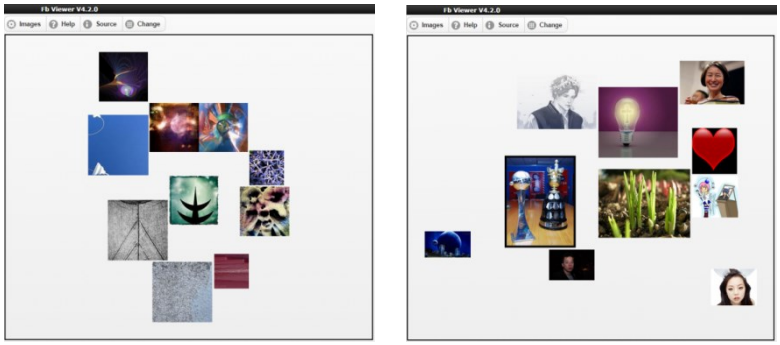


Figure B.12 - Feedback summary screensfor abstract (left) and emotive (right)images. The collages are interactive in that tapping an individual image opens a full view of the image.

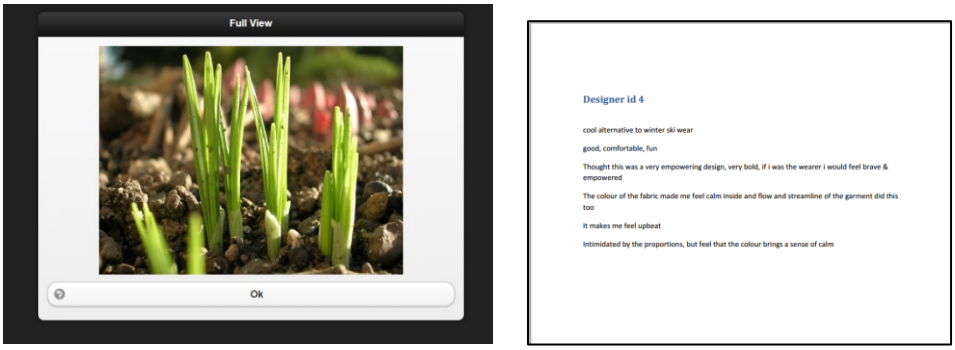


Figure B.13 - Image full view screen (left) and text list screen (right).

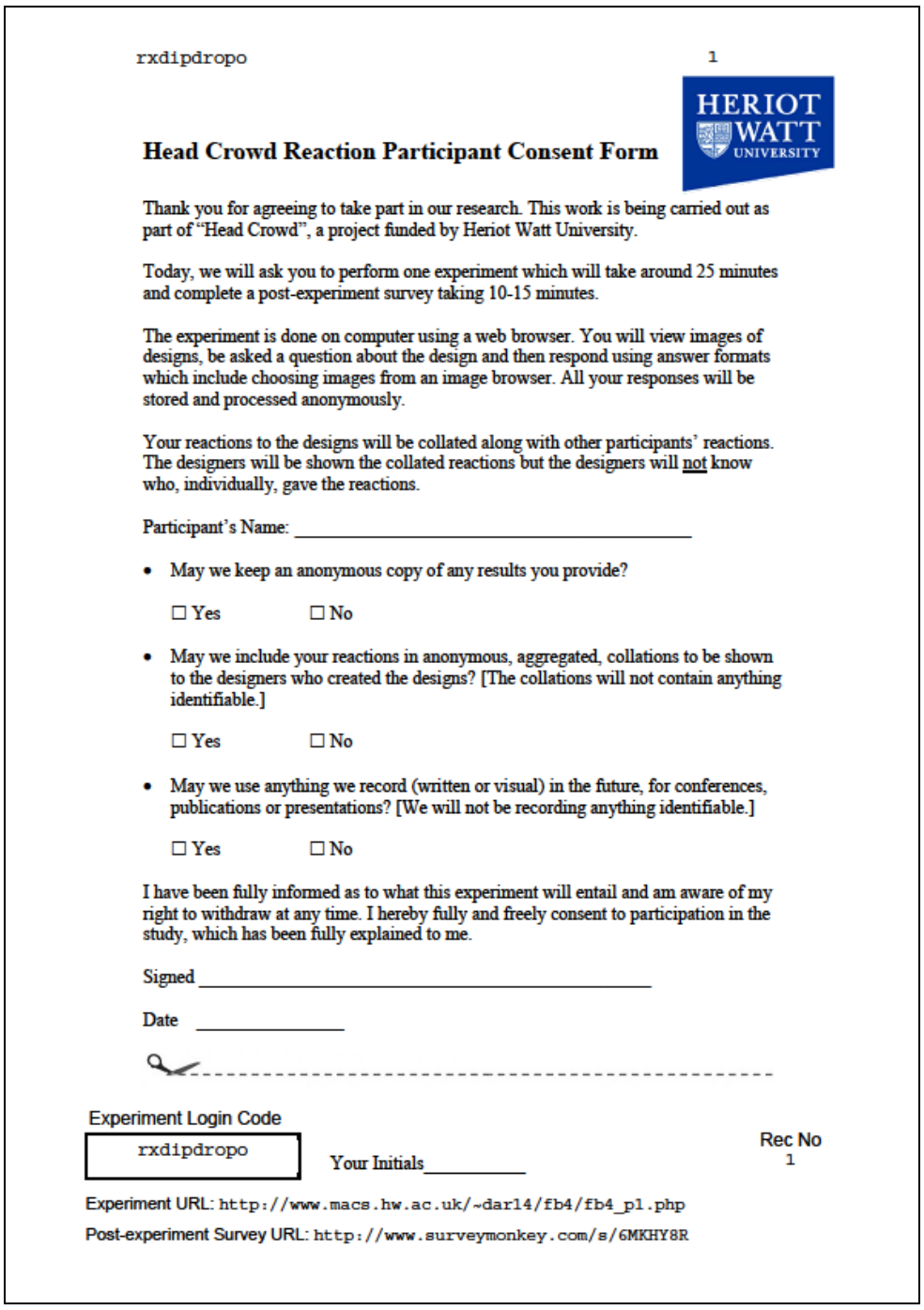
Appendix C Main Evaluation Study

This Appendix accompanies Chapter 10.

Additional Material

The Additional Material CD contains directories relating to the Main Evaluation Study including code, input and output files and documentation.

Feedback Participant Consent Form



rxdipdropo 1

Head Crowd Reaction Participant Consent Form

Thank you for agreeing to take part in our research. This work is being carried out as part of "Head Crowd", a project funded by Heriot Watt University.

Today, we will ask you to perform one experiment which will take around 25 minutes and complete a post-experiment survey taking 10-15 minutes.

The experiment is done on computer using a web browser. You will view images of designs, be asked a question about the design and then respond using answer formats which include choosing images from an image browser. All your responses will be stored and processed anonymously.

Your reactions to the designs will be collated along with other participants' reactions. The designers will be shown the collated reactions but the designers will not know who, individually, gave the reactions.

Participant's Name: _____

- May we keep an anonymous copy of any results you provide?

  ☐ Yes ☐ No

- May we include your reactions in anonymous, aggregated, collations to be shown to the designers who created the designs? [The collations will not contain anything identifiable.]

  ☐ Yes ☐ No

- May we use anything we record (written or visual) in the future, for conferences, publications or presentations? [We will not be recording anything identifiable.]

  ☐ Yes ☐ No

I have been fully informed as to what this experiment will entail and am aware of my right to withdraw at any time. I hereby fully and freely consent to participation in the study, which has been fully explained to me.

Signed _____

Date _____

Experiment Login Code

rxdipdropo

Your Initials_____

Rec No 1

Experiment URL: http://www.macs.hw.ac.uk/~dar14/fb4/fb4_p1.php

Post-experiment Survey URL: http://www.surveymonkey.com/s/6MKHY8R

Each form carried a unique task login code. This was to allow task responses of individuals to be anonymously collated. It also allowed the post-task survey responses to be tied to the task responses. (Participants were asked to enter the code and rec.no on the post-task survey). Thus the entirety of each participant's input to the study could be collated and anonymously attributed as that of one individual. The participants read and signed the form, detached the login slip and handed in the form. Participants retained the slip for reference when logging into the task application. They were asked to write their initials on the slip to help guard against mix-ups (e.g. another participant using their login) in case they laid the slip down near another participant when starting the task. Feedback participants were termed "Reaction Participants" on the form so as to avoid them thinking of the task as giving feedback in the conventional sense but to help them focus on emotions when answering the question "How did the design make you feel?".

Screens from the Feedback Task Application

Below is a selection of screens from the feedback task in the main evaluation. It gives the 'flavour' of the interface and depicts two of the important stages. However, a full sequence of screens illustrating a unit of work in the task can be found in the Additional Material "Main Evaluation Study" folder.

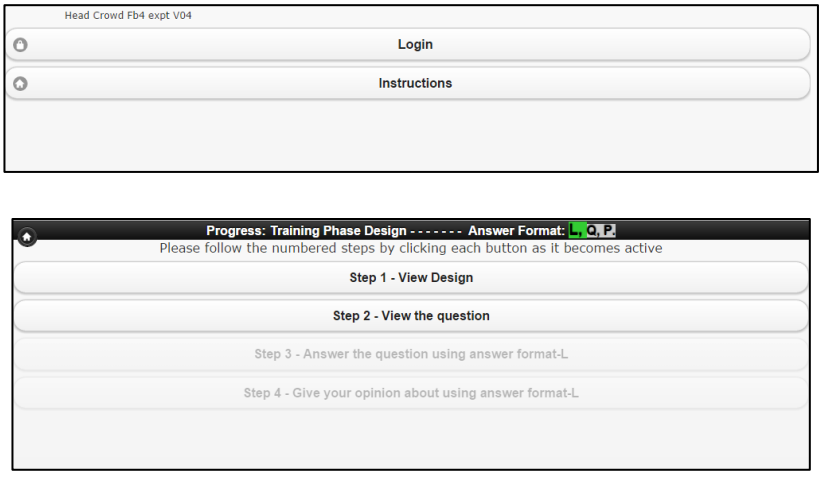


Figure C.1 - Interface screen: Start screen.(top) and Stepping through the task (bottom).

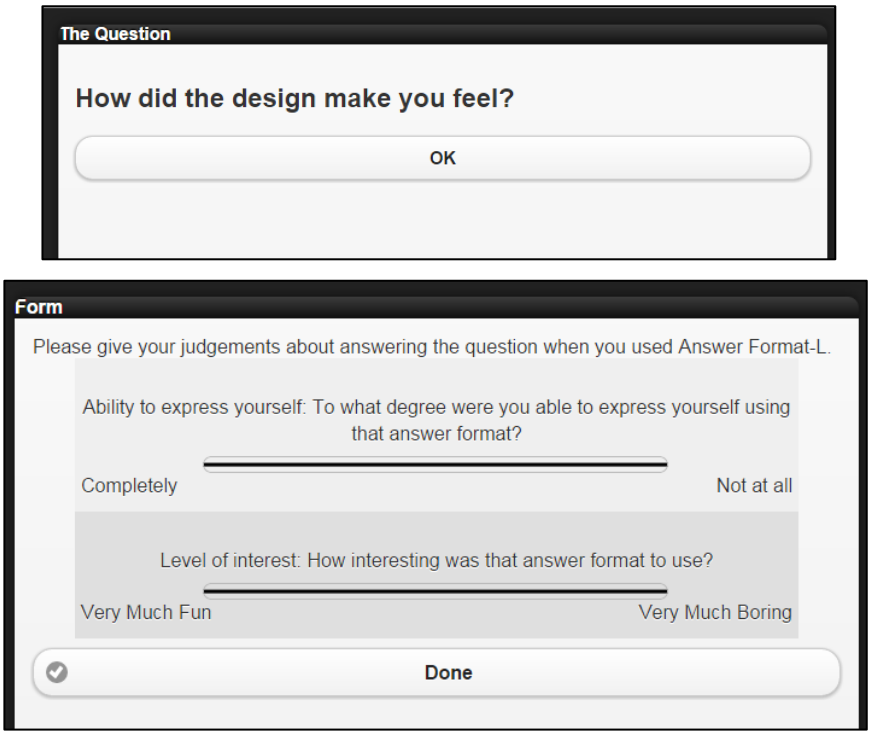


Figure C.2 - Interface screens: Viewing the question (top); VAS items prior to being set (bottom).

Post-Task Survey

Feedback participants completed a survey after the feedback task. (One participant failed to complete the survey).

The purpose of the questions fell into these categories:

1) Participant ID: fields to allow the survey answers of each participant to be matched anonymously to their task data;
2) Establishing whether or not the participant had understood what they were doing in the task;
3) Seeking opinions about the visual feedback formats;
4) Providing an opportunity for open-ended comment; and
5) To ask participants to report on the issue of "freedom of expression" because, from the pilot, it had been decided to discard the VAS item measuring this during the task (see 9.10.2).

The two tables below detail the questions in the survey showing the response types. The category (listed above) is noted for each question to indicate its purpose.

Q. No.	Wording	Response type	Cate-gory
Page 1: How the Experiment worked for you: Any answers you give during this post-experiment survey are purely for analysis by the research team. The designers whose designs you viewed will not see them. This page is mainly about finding out if the experiment application worked for you.			
1	Please enter the login code you used to login to the experiment	Text field	1
2	Please enter the number (Rec No) printed on the right of your experiment login slip (It should be a 1 or 2 digit number	Number field	1
3	Did you understand all the questions in the experiment? Supplementary: Please describe anything you did not understand.	Yes/No Text field	2
4	Did the experiment go smoothly for you? Supplementary: Please describe any difficulty you encountered.	Yes/No Text field	2
5	What type of computer did you use to do the experiment?	Options + text field	2

Table C.1 - Page 1 of feedback task post-task survey showing question wordings, response types and question categories.

Q. No.	Wording	Response type	Cate-gory
Page2: Seeking your views and thoughts about the answer formats On this page you may need to refer to specific answer formats. To avoid any mix-ups, instead of labelling them Q, P and L. Please use the following labels for the three formats. The text format: "Text" The images which featured facial expressions and people: "Emotion images" The small images of textures and abstract views: "Abstract images".			
6	Please take a minute to think about whether or not you held back in some of your answers during the experiment. Perhaps at times you toned down your reaction to some of the designs so as not to hurt the feelings of the designers or so as not to appear too harsh? Perhaps at times you felt no inhibitions? Did the degree of freedom you felt vary between the three different answer formats? Please describe your thoughts on this referring to the three answer formats (using the labels: "Text"; "Emotion images"; "Abstract images")	Text field	5
7	Please rank the answer formats in order of your overall preference: Text Abstract images Emotion images	Drop-down list or drag-drop	3
8	Please try to describe the reasons for the ranking you gave to the formats in your answer to the previous question.	Text field	3
9	When looking for images to express your answers, how easy or difficult did you find the image browsers to use? Very Difficult; Difficult; Neither Easy nor Difficult; Easy; Very Easy	5-point Likert	3
10	Please tell us what you think of the idea of communicating about designs using images versus text when using technology such as computers, tablets, and smart-phones.	Text field	3
Page 3: And finally A last opportunity to comment			
11	Is there anything else you wish to add? If so, please use this box.	Text field	4
	Thank you for taking part in the Head Crowd research		

Table C.2 - The remainder of post-task survey showing question wordings, response types and question categories.

Detailed Feedback Task Results

The medians of the five experiment phase readings over the six measures (three formats by two VAS items) for each of the 32 feedback participants are shown below. Also shown (in the column headed "Preference group") are each participant's first feedback format preference as stated in their post-task survey response: 0 = text; 1 = an image format; 2 = survey not completed.

S/No	Preference group	Interest Text	Interest Emotives	Interest Abstracts	Utility Text	Utility Emotives	Utility Abstracts
1	1	287	299	268	310	241	276
2	1	214	165	149	149	192	180
3	1	96	230	15	119	245	61
4	0	276	184	123	0	341	245
5	0	333	310	272	169	280	211
6	1	253	272	268	260	287	268
7	1	253	165	50	31	57	27
8	0	280	268	272	314	329	337
9	1	8	27	11	4	4	11
10	1	230	176	203	234	184	199
11	1	241	356	341	172	326	349
12	0	46	214	142	23	195	130
13	1	107	57	107	126	123	149
14	0	80	257	287	61	188	345
15	0	249	188	169	73	276	237
16	1	199	134	100	184	188	123
17	1	42	34	54	42	31	50
18	1	149	195	123	8	115	103
19	0	195	195	188	15	138	161
20	1	299	130	50	123	260	134
21	1	352	73	61	46	184	46
22	1	195	107	119	134	115	123
23	0	203	283	218	130	306	276
24	1	149	119	80	100	107	119
25	1	260	264	241	257	249	234
26	1	31	42	34	38	27	27
27	2	100	100	100	103	103	100
28	0	241	199	192	65	142	192
29	0	184	379	379	31	379	379
30	0	264	283	303	188	199	218
31	1	214	234	230	241	253	253
32	1	192	184	192	92	192	176

Table C.3 - Detailed results from the Feedback task.

K-S Tests for Normality

All six score distributions were tested for normality using the K-S test (Field 2009 p144). See table below. For all 6 distributions the significance value is not less than

0.05 indicating that none of them deviate significantly from normality (Field, 2009 p.246).

Distribution	K-S test statistic	df	Sig.(*p*)	Passes test?
Utility				
Text	0.10	32	0.20*	Yes
Emotives	0.11	32	0.20*	Yes
Abstracts	0.08	32	0.20*	Yes
Interest				
Text	0.14	32	0.08	Yes
Emotives	0.08	32	0.20*	Yes
Abstracts	0.11	32	0.20*	Yes

Table C.4 - K-S tests for the six results distributions0.20 indicates that 0.20 is the lower bound of the true significance.*

The KS test was done through the SPSS explore menu; the Lillifors Significance correction was applied (Field, 2009 p147).

Results from Post-Task Survey Question No. 4

Themes from question no. 4 from the survey are shown in the table below as the "freedom of expression" issue was one of specific interest in the study.

Theme Description	
Sub-Theme	Number of responses
Quoted responses	
Not holding back	
Not holding back irrespective of format	2
"was completely honest and did not hold back with regard to the way in which I answered the questions using Text, Emotional Images and Abstract Images." *"I did not hold back or change try to tone down my opinion for anything"*	
Holding back	
When using **Text**	5
"I also think I held back in the text answers as I found it hard to put my feelings into words."	
"Text: I felt I held back slightly with the words as words are more obvious and at times hurtful."	
"Text- I felt I had to tone down my views as to not offend the designer"	
"I did hold back slightly, i did not want to offend the designer however I did like everything which i was shown."	
Also by implication: *"Emotion images/ Abstract Images- I did not hold back"*	
When using **Emotive images**	1
"Emotion Images: It was easy to express with these pictures but i felt it wouldve been harsh to choose some of the boredom images"	
Abstract images not hurting feelings	
-	1
"I think I found the abstract images the easiest to use as I wasn't necessarily hurting	

anyone's feelings by not liking their design."	
Not addressing the issue in the question; addressing other issues instead	
Most were off-topic but useful	22; three examples quoted
"text - was easier to explain what I feel emotion - doesn't give the whole idea of what i fell abstract images - was more fun to use it" *"I preferred the use of text as I was able to use my own words to say how i felt about a design. The emotional images worked ok, though I often found it difficult to find an images that perfectly reflected my emotion. abstract images i didn't fully understand and felt i was selecting any image slightly relevant to my emotions"* *"Text - fair Emotion and abstract images – instincts"*	

Table C.5 - Post-task survey: themes from Q4 on the issue of "freedom of expression".

The Decision Not to Discard Feedback from the Training Phase

As is described in 10.1.1 it was wished to maximise the amount of feedback to be shown to each designer participant. Rather than discard the feedback responses from the training unit, the text responses were examined to ensure that there were none that conveyed the impression of them being formed carelessly by feedback participants as they were provided during the training phase. It was clear from the text feedback that the participants had given genuine feedback during the training phase. This also indicated that the image feedback would also be genuine. It might be argued that feedback participants would be unfamiliar with the full extent of both image sets during the training phase. Indeed this would be true. In fact it was expected that, for all feedback participants, learning about that the browsers, would have continued throughout much of the task, not just the training phase. The effect of this might well mean that were a participant to view the same design again they might make different image choices in the light of greater familiarity with the image sets. This does not, however, render "inexperienced" image feedback invalid as such choices were still made in response to the stimulus design and question. Also, as designs were presented in a random order the image feedback corpus would contain a balance of "inexperienced" and "experienced" feedback.

Generating the Feedback Image Summaries

The feedback text and image selections for each designer were collated by running queries on the feedback task database. These produced two image selection lists (ISLs) in CSV files for each designer. The ISL files were the input, along with the respective image set perceptual data and 3D non-metric MDS coordinates files for the Abstract500 and Emotive204 image sets, to the MATLAB scripts for image summarisation. The perceptual data for the Abstract500 was its associated similarity matrix. The perceptual

data for the Emotive204 were the emotion tag vectors for the 204 images and their accompanying labels file.

Inputs to "myKmeansOnlyForFb4Emotives.m" to make the emotive summaries
12 ISL files e.g. "E1.csv" D1's emotive image feedback selection list
simFileName='eciQc3pt1FullE2kMTTOFXDExRjctdBalByTermTop13fltrV3At35Pop204-LocnVectors.csv'
labelsFileName='eciQc3pt1FullE2kMTTOFXDExRjctdBalByTermTop13fltrV3At35Pop204-LocnVectorsLabelsIDX.csv'
coordsFileName='eciQc3pt1FullE2kMTTOFXDExRjctdFV3At35Pop204LocnVsDmat-cityblock3dMdsCoords.csv';
Outputs were 12 files on for each designer e.g. "E1collatedClusterOnlyInfo.csv" for D1's emotive summary.
Inputs to "myKmeansOnlyForFb4Abstracts.m" to make the abstract summaries
12 ISL files e.g. "A1.csv" D1's abstract image feedback selection list
simFileName='abstract500augSimFeb2012mturk.txt'
The labels are implicit in the Abstratc500 similarity matrix
coordsFileName='Abstract500_3d_mds_coords-NONMETRIC.csv';
Outputs were 12 files on for each designer e.g. "A1collatedClusterOnlyInfo.csv" for D1's abstract summary.

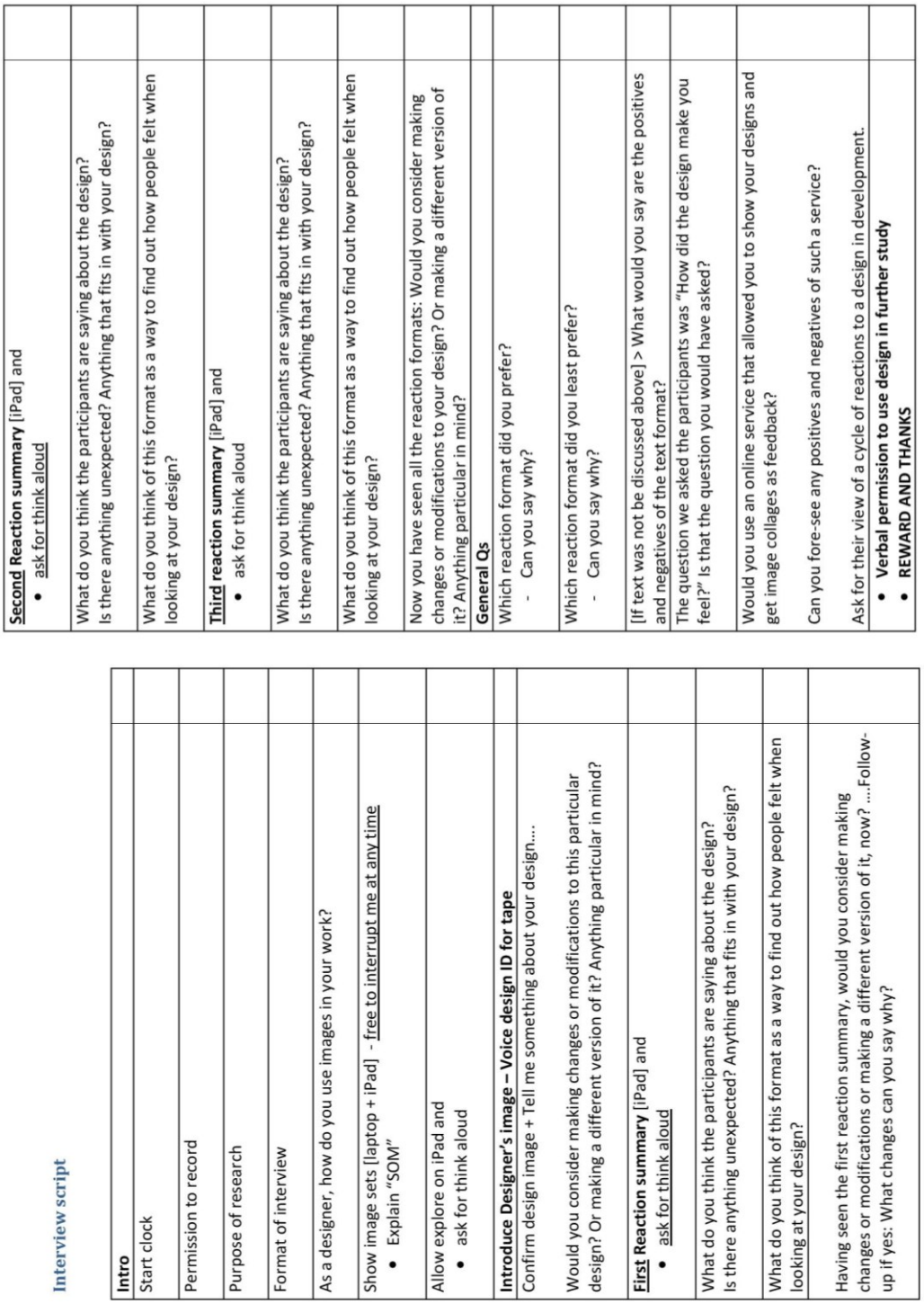
Table C.6 - Script names, input file names and output filenames. See Additional Material CD for files.

Record of the Randomised Format Order in Designer Interviews

Designer Participant	Format order	Designer Participant	Format order
1	A-T-E	7	E-A-T
2	E-T-A	8	A-E-T
3	T-A-E	9	A-E-T
4	T-E-A	10	E-A-T
5	T-E-A	11	A-E-T
6	E-A-T	12	A-E-T

Table C.7 - Main evaluation interviews: record of format presentation order. Formats A, E, and T were abstract images, emotive images, and text respectively.

Designer Interview Script



Interview script

Intro	
Start clock	
Permission to record	
Purpose of research	
Format of interview	
As a designer, how do you use images in your work?	
Show image sets [laptop + iPad] - free to interrupt me at any time • Explain "SOM"	
Allow explore on iPad and • ask for think aloud	
Introduce Designer's image – Voice design ID for tape	
Confirm design image + Tell me something about your design.... Would you consider making changes or modifications to this particular design? Or making a different version of it? Anything particular in mind?	
First Reaction summary [iPad] and • ask for think aloud	
What do you think the participants are saying about the design? Is there anything unexpected? Anything that fits in with your design?	
What do you think of this format as a way to find out how people felt when looking at your design?	
Having seen the first reaction summary, would you consider making changes or modifications or making a different version of it, now? ....Follow-up if yes: What changes can you say why?	

Second Reaction summary [iPad] and • ask for think aloud	
What do you think the participants are saying about the design? Is there anything unexpected? Anything that fits in with your design?	
What do you think of this format as a way to find out how people felt when looking at your design?	
Third reaction summary [iPad] and • ask for think aloud	
What do you think the participants are saying about the design? Is there anything unexpected? Anything that fits in with your design?	
What do you think of this format as a way to find out how people felt when looking at your design?	
Now you have seen all the reaction formats: Would you consider making changes or modifications to your design? Or making a different version of it? Anything particular in mind?	
General Qs	
Which reaction format did you prefer? - Can you say why?	
Which reaction format did you least prefer? - Can you say why?	
[If text was not be discussed above] > What would you say are the positives and negatives of the text format?	
The question we asked the participants was "How did the design make you feel?" Is that the question you would have asked?	
Would you use an online service that allowed you to show your designs and get image collages as feedback? Can you fore-see any positives and negatives of such a service?	
Ask for their view of a cycle of reactions to a design in development. • **Verbal permission to use design in further study** • **REWARD AND THANKS**	

Interview Findings

In the sub-sections below the main themes arising out of the interviews are described along with what the interview evidence leads us to conclude about them.

Theme 1 Interpreting the Feedback

While viewing and exploring a visual feedback summary, designer participants would develop their interpretation of the feedback. Here while viewing emotive image feedback on her design for a bar interior and successively expanding the individual component images: *"Mmm. I think they are talking about the mood in this one. How, like, people here, socialising; they are happy. Something crazy going on here [little laugh]. And, [I] don't really understand this one here. Like you can just sit down by yourself and get lost in your thoughts. They are talking about the mood here, I think."*[D5].

A similar process seemed to occur with ambiguity in the text feedback with the designers assigning a message or messages to comments and groups of similar comments e.g.: "*[quoting from her text feedback]"planning and organising, sense of group". Yeah, 'cause it's sort of the way that the chairs are laid out and stuff."*[D12].

One example of meaning not being discerned in an image in the emotive feedback but a message about colour still being assigned: "*When I look at the lego hands, it's got basically all the colours that I've used. [and later] I didn't really understand what the hands meant, but the colours I understand*."[D3].

The designers addressed ambiguity in the images assigning a message to an image or group of images on a summary.

Theme 2 Inspiration to Make Changes

Sometimes a designer participant was immediately inspired to make a specific change to their design. Here after viewing her abstract image feedback summary: "*I was looking at and thinking that was earthy and very cold, it is not the environment I really wanted. So yes, it is making me think, definite change of textures, if that is how they see it as cold and mechanical. I didn't think that would be the reaction you would get but that is good though. Good feedback*" [D11].

Sometimes a less specific change was motivated. Here after viewing emotive image feedback*: "I'd make it a nicer visual. I'd make… I'd refine it a bit more. I'd put more detail into it. I think. [be]cause it [her design image] is a bit boring."* [D6].

These are two examples of designers finding motivation for design changes in the visual feedback. The abstract image feedback was being read for colour and texture ideas

while the emotive image feedback was prompting a change due to the designer reading the emotion, boredom, in the feedback. Here the emotion being read was negative, and this is discussed in another theme below, "*Negative feedback*".

Theme 2.1 Inspiration – A Quantitative Analysis

The table below details which designer participants indicated inspiration from their first feedback. See also Table C.7 which details the random order in which each participant was shown their feedback.

	Text first			**Abstracts first**					**Emotives first**			
Designer Participant	4	3	5	11	1	8	9	12	10	7	2	6
Inspiration?	X	X	X	✓	X	*	✓	X	✓	X	✓	✓
Totalsa	0/3			2/4					3/4			
Total participants asked about inspiration after first feedback									11			

*Table C.8 - Quantitative analysis of inspiration after first feedback. Explanation of ✓ and X are below in the text. * NB: D8 was not asked about inspiration this way as other themes were pursued early in that interview.*

A ✓ symbol means answers ranging from: *"I think I feel I should maybe [be] more natural considering I am trying to give it a warmth feel."[D1];* and also: *"Yeah. Maybe to add some more interest to it."[D2];* and including: "Believe it or not, yes because I was looking at and thinking … Good feedback"[D11] (See the rest of D11's quote above in Theme 2); and then later "*Yes definitely because that's what gave me the ideas of what I could improve on. Definitely*."*[D11]*; to: an immediate, *"Absolutely yes!..."[D9].*

An X means anything from. *"Nope"[D12]*; and including: *"Not really because most of the, well if I got loads of negatives, that would be a different story…"[D4]*; to: *"Em. [pauses thinking]. Em I'm not sure. Em. If I was going further with the design I would like… would use that to do that… I can't think of anything right now."[D7].*

Theme 3 Abstract Image Summaries as Mood Boards

The abstract image summaries were likened to mood boards. While talking about her abstract feedback summary:*"...Just sort of represents what is actually there [in her design]…because it is outdoor there is a lot of green and a lot of wood…Yes the look is similar to what my mood board would look like before it."* [D12].

D12 continued on this theme later when suggesting that she would use the abstract image feedback as a presentational tool for describing the design to others including those who commissioned the design:

"... to show like if it was a presentation and you were then saying, "Well, I've actually surveyed all these people and this is what they thought of it", and then to show that [indicating the abstract summary] and that being similar to what I had done at first [referring to her own mood board made at the outset of the design project] ... like we have to stand and present all our work every time that we finish it. So then to stand and present, and to say I've surveyed, or people have surveyed 15 people and that's what came back." [D12].

Here two uses of the abstract image feedback are indicated. Firstly it could act as a form of reverse-engineered mood board confirming that the designer's originally planned "mood" for the design was being communicated as intended. We saw this operating in the negative when D11 (quoted in Theme 2) was motivated to make a change by her abstract feedback because she was responding to her mood-board-style reading of that visual feedback. The second suggested use here for the abstract feedback, is when discussing the design with others, such as a client, to demonstrate the mood actually conveyed by design.

Theme 4 - Negative Feedback

This theme merited division into sub-themes:

Theme 4.1- Perception of Negative Feedback Across Formats

Negative feedback was a topic arising in discussion from participants while viewing text feedback and emotive image feedback. However it was not mentioned by any participant while viewing the abstract image feedback.

Combining this with the observation that changes could be motivated by the abstract feedback (see Theme 2) suggests that the abstract feedback can be inspirational without being perceived as threatening.

Theme 4.2 The Tendency to Focus on Negative Feedback

The tendency for participants to focus on negative feedback is demonstrated by the following quantitative analysis of each participant's text feedback and how they chose to scan it during interview. Each participant was asked to say what they were thinking

as they viewed each of their feedback formats. When they viewed the text feedback (a simple list in random order), typically they would scan down the list and read out loud several of the comments and describe their interpretation of them. Only three of the 12 participants read out the first item on the list. Nine participants skipped one or more items to read out another that they had focussed on first. Eight of those skipping comments chose to focus on a negative comment first, while only one skipped to a positive comment (A negative comment was defined as a comment with a clear negative element in it. A "positive" comment was defined as any comment not defined as negative, and so included neutral comments. The mean percentage of negative comments in the 12 participant's text feedback was 30.1%; SD 20.4%; Median 24.3%).

This was acknowledged in discussion. One participant when asked why she had stopped at a specific comment: "*Just 'cause the first two sounded quite positive. [laughs]… I was enjoying reading it up to there [laughs].*" [D7]. Another participant when it was pointed out that the list contained more positive comments than negative: *"You just can't help but read the bad stuff".* [D6]. Also, when talking about the text format in general: *"There's lots of nice comments on here though. I'm just picking out all the bad ones."* [D2]

Negative feedback was also perceived in emotive image feedback summaries. Participant D3's emotive image feedback summary contained only one negative image out ten. (The image was of a man covering his eyes with his hand). The size of the images on the summaries varied with the population of the feedback response cluster they represented, but the single negative image that D3 chose to focus on only represented just 20% of the total area covered by all ten images on the summary. D3's words are quoted in Theme 4.3 below as they also pertain to that theme.

One interpretation of this focus on negative feedback over positive is that the designers were valuing the negative feedback over the positive however unpalatable it might have been for them.

Theme 4.3 - The Impact of Negative Text Compared to Negative Emotive Images.

One participant felt that the negative feedback received via the emotive images was more impactful than text feedback. Here she is referring to the single negative image in the summary: *"I think the emotive images are quite hard to look at because it is peoples' emotions towards your, em, design. And if an image is that big, it does kind of pull you back and like, "Why?". But that might be because they don't understand*

Gaudi[4]*, or as I say, the design was quite cramped, and that's probably why they didn't understand it as well… when you look at the images, they'll be stuck to you. Whereas the writing it doesn't really stick much to you. You just read it and you're like "Ok." But the images, you're like "Wow!" It's almost like you can see that person's emotion…when they are picking this image."* [D3].

For another participant negative feedback via text was more impactful than the emotive image feedback: *"Looking at that [the emotive image summary]. I'd say I'm more relaxed looking at the images than the text. I'd say I'm more relaxed looking at them. Even though I've read this [the text list] and this dude's bored and this wee girl's bored and that guy's confused [pointing to component images in the emotive summary]. It's just less threatening than the text. 'Cause people have a way of…people have a way of putting things that might not be effective to whoever's getting criticised. Em. So the images is a good idea in that way."* [D6].

While there was disagreement within the designers on whether negative feedback had more impact as text or as emotive images this does demonstrate that the designers were able to get negative feedback via the emotive image format, and because they showed a keen interest in negative feedback this would indicate that the emotive image feedback would be of value to them.

Theme 5- Effectiveness at Finding Out How People Felt

When asked how well the text feedback answered the question "How did the design make you feel?", here D6 points out that the text comments had actually strayed into a critique rather than talking about feelings: *"[quoting from the text feedback] "modern, young, cool, stylish, good interior for shoe display". I think a lot of them have got the gist of it, because the flaws that they pointed out, I would also point out as well. Like the fact that it's not that big and it's a bit busy and stuff like that [referring to her design]. Em. But yeah. No-one's really said how they feel really. Well, [quoting again] "I felt uninspired" There's one. But that's it... Yeah. I think the emotive images work better than the text…[be]cause it's fair enough if they were critiquing it, but they're not. They're meant to be saying how they feel and no-one's really [done that]."* [D6]. Another participant: "*what they said in the text isn't exactly feelings*" [D8].

[4] Participant D3's design included an homage to the Spanish architect, Gaudi.

Those designers clearly think that the emotive images have allowed those giving the feedback to focus on communicating their emotions more effectively than when using the text format.

Another participant on the effectiveness of images for emotion: *"I like that [emotive image summary]. 'Cause it shows emotion as well, yes, mostly like emotions that what people would feel…It's a good way of getting their understanding."* [D2].

Theme 6 – A Service Offering the Visual Feedback

This theme merited division into sub-themes:

Theme 6.1 – Would Designers Use the Visual Feedback Service?

After viewing and discussing the feedback formats participants were asked if they would use an Internet service which allowed them to upload a design and receive feedback in the *visual* formats. Ten of the designers answered emphatically in the positive, one was neutral and one (D12) initially wished for text feedback but moved on to develop the idea of using the abstract feedback as a presentation tool. One participant was particularly effusive: "*I'd love that! I'd absolutely love that yeah!"* [D8].

From this it is clear that the designer participants valued the visual feedback formats and wanted more.

Theme 6.2 – Present Prototypes and Refine through Cycles of Visual Feedback

The designers were probed on how they would use the service. Specifically, would they present a prototype or finished design? If they presented a prototype would they respond by changes and seek further feedback? The participants were unanimous in the view that presenting a prototype and developing it in response to crowd feedback would be the way to use the service; e.g. when asked if she would put up the finished design for feedback D3 responded: "*No. It would actually be much easier if I did it during the process. So it would be easier to get a better final product. Rather than putting the final product…"*

Detailed results for the designer participant format preferences

Designer ID	Text	Emotive images	Abstract images
1	3	2	1
2	1	2	3
3	2	3	1

4	1	2	3
5	1	2	3
6	3	2	1
7	2	3	1
8	3	1	2
9	3	1	2
10	2	3	1
11	1	3	2
12	1	3	2

Table C.9 - Detailed results for the designer participant preferences. "1" means that format was the participants most preferred format, "3", least preferred.

Detailed Analysis of Reasons for Designer Participant Format Preferences

The reasons given by the designer participants for why they ranked a given feedback format first were analysed. Those reasons are described in the below with supporting quotes. The themes from these reasons are summarised in the table below.

Reasons for ranking Text first
Unexpected depth in some comments; also how the text says how they feel
Because some things I can understand but some go really into depth over just one image[her design] – it is their thought process which is really not what I was thinking of at all when I was designing it which is quite interesting. I like how the text says how they feel as well – like angriness and stuff.[D11]
Easier to understand how people felt by text. You might think about what feelings the images meant but misunderstand the intended meaning
Designer12 - Its just easier to understand. Like although the first one was like easy enough to understand how the people felt. I think that would have come across a lot better with text. It's easier to understand. *Researcher – So you are getting more detail.* *Designer12 - Yes. Definitely. And its not like, its again like you could sit and guess what people are meaning with that and you might not actually get what they are actually were thinking when they were selected it.[D12]*
Close decision between Text and Emotives; Text is the most honest
I think [thinking] the text… *Researcher – Uh huh?* *Designer2 – ...I think yeah. It's between the text or the emotion one, I feel would be most helpful.* *Researcher – Eh, so choose.* *Designer2 – [laughs] Eh [thinking]. Text.* *Researcher – Ok. And can you say why you'd be choosing text?* *Designer2 –'Cause I think it's the most honest.[D2]*
An image can be ambiguous due to one person focussing on one part of it while another focuses on a different part which has a different meaning.
It[text] is easier to understand and it is what it is. Where I am looking at the image, the person …sport running with the ball. I might be seeing someone running with the ball but someone else might be looking at the t-shirt and the t-shirt could be going back to my design.[D4]
Text is clear
it's clear. It's all clear. You don't have to make sense of it.[D5]
Reasons for ranking Abstracts first
Visual person; it is interesting; How the crowd have grasped the forms, colours and textures from the design
I think just 'cos I am a visual person so to see people's feedback visually is like interesting.

Like how they have linked textures and things and like even the colours. Yes, the circular forms and stuff how they have kind of grasped that?[D10]
Text not as interesting; images you look more into; text is just text
... it is just because it [text] is not as interesting. You would look more into that, whereas someone's text, well that's just it – that's what they think… [D1]
Abstract is a replica or distillation of the design; The abstract images could mean anything but the emotive images clearly mean something (and that might be negative)
Designer3 – I prefer the abstract one because it is basically a replicat of my design, but in photos. *Researcher – Ok.* *Designer3 – That's why I prefer that one.* *Researcher – And so, when you say it's a replica of your design, in terms of your preference, why do you think it is that that's making you like it, over the others.* *Designer3 – Because, it's like if you took a telescope to my design you would see these shapes, you would see these curves, you would see these colours, you would see these lights, and I think that's why I like that abstract one 'cause it really does replicate it. It's almost like my design has been pulled apart…* *Researcher – Uh huh?* *Designer3 – ... and been zoomed into, so you can see all this. I think that's why I like the abstract one.* *Researcher –and…* *Designer3 – 'cause you can see these colours, you can see these curves.* *Researcher –So, in terms of when people have looked at your design then, that's, eh… and they're coming up with what you think is a distillation of it…* *Designer3 – Yeah.* *Researcher – … Yeah?* *Designer3 – A very good distillation. [little laugh]* *LATER* *Designer3 – I would only ask them for abstract images [laughs].* *Researcher – So you'd be saying give me your feedback using that image set? [indicating abstract image set]. Right ok.* *Designer3 – Cause the abstract images can mean anything*. Whereas the emotive images they obviously mean what they mean.[laughs][D3]
Abstracts had a "happier" impression compared to the emotive images
'Cause it seems nicer than the other ones. *... And also 'cause these pictures seem a bit happier that those ones [little laugh] [indicates the emotive collage].* *... [Laughs] They are happy colours.[D6]*
Reflected the design; The abstract images were more understandable than the emotive ones.
Just because how it turned out it reminded me of the image. *Researcher – Your design?* *Designer7 – Yes.* *Researcher – Ok.* *Designer7 – And I think I can understand like the shape and the colour more in that one [abstract collage] than the likes of the emotive one.[D7]*
Reasons for ranking Emotives first
Emotives give more understanding; Text is too conventional; Emotive images allow you to take what you want from it. If you are sensitive you can take out good things; However emotives can still show negative opinions.
I think that one's[Emotives] more understanding. The one with the emotive ones. I really like the text but that goes back to like how its always done? *Researcher- Yes.* *Designer8 – You know it's always, em, people's opinions are always put across by text and whether that's a good thing or a bad thing I don't know because it gets your point across in a very direct manner?* *Researcher- Yes*

Designer8 – Whereas maybe from these images especially the emotive ones, you can take out of it what you want a bit more? So maybe if you are a bit sensitive about your design you could take out the good things but then as I said you know. You need peoples bad opinions to kind of improve it so you need an overall thing. But, you know even so, that [indicating the emotives collage. kind of still puts across peoples maybe not-so-good opinions. But I think the abstract one was really interesting but emm,[D8]
The emotive image made the crowd think about what they thought of the design. The emotive images gave a different perspective on the design.
[pauses thinking]. Eh. Oh god. I'm gonna say the emotive…. *'cos I initially didn't get some of the images instantly. …* *So it kind of made me think that the people that have looked at this image have kind of thought about it…* *.. and gone… well that's, this is what it's made them think of.* *Whereas I maybe wouldn't have though that.* *…* *So it gives me another sort of perspective on it.[D9]*

Table C.10 - Designer participant reasons for ranking a given format first with supporting quotes.

Summary of the Reasons for Designer Participant Format Preferences

Reasons for ranking Text first	**Themes**
Unexpected depth in some comments; How the text says how they feel. [D11]	T1,T2
Easier to understand how people felt by text. You might think about what feelings the images meant but misunderstand the intended meaning. [D12]	T2,T4
Close decision between Text and Emotives; Text is the most honest. [D2]	T3
An image can be ambiguous due to one person focussing on one part of it while another focuses on a different part which has a different meaning. [D4]	T4
Text is clear.[D5]	T2,T4
Reasons for ranking Abstracts first	
Participant stated they are a visual person; It is interesting; how the crowd have grasped the forms, colours and textures from the design. [D10]	A1, A2
Text is not as interesting. Images you look more into. Text is just text. [D1]	A3
Abstract is a replica or distillation of the design; The abstract images could mean anything but the emotive images clearly mean something (and that might be negative). So Abstracts allow one to avoid negative feedback. [D3]	A2, A4, A5,
Abstracts had a "happier" impression compared to the emotive images. [D6]	A5
Abstracts reflected the design; The abstract images were more understandable than the emotive ones. [D7]	A2, A6
Reasons for ranking Emotives first	
Emotives give more understanding; Text is too conventional; Emotive images allow you to take what you want from it. If you are sensitive you can take out good things; However emotives can still show negative opinions. [D8]	E1, E2, E3, E4
The emotive image made the crowd think about what they thought of the design. The emotive images gave a different perspective on the design. [D9]	E5, E6

Table C.11 - Summary of designer participants' reasons for ranking a given format first. The themes refer to Table 10.5 (p.162) summarising the themes.

Comparing the Pilot Feedback Task Results with the Main Study

This subsection compares the VAS item results from the pilot with the main study. There was a noticeable difference, i.e. the VAS readings were generally more positive in the pilot. Indeed they were so positively skewed that they required log transformation to fit a normal distribution. (The score distributions concerned are illustrated in Figure 9.3, Figure 9.5, Figure 10.4, and Figure 10.5). Aside from the size of the participant groups (10 for the pilot and 32 for the main), the areas of difference between the pilot and the main study are set out in the table below.

Aspect of conditions	Pilot	Main study
Task materials	Paper task sheet prompting use of computer interface and with paper recording of VAS responses	Fully integrated computer interface leading the task and recording responses
Task workflow	Fewer designs (4) and fewer VAS readings. Less repetitive.	More designs (6) and more VAS readings. More repetitive.
Task duration	25 minutes (median: 23; SD: 0.5; max.: 44; min.: 17). Longer time on task.	19 minutes (median: 18; SD: 5.8; max.: 35; min.: 10). Shorter time on task.
Physical location and setting	A quiet corner of an open plan garment workshop where the participants were already working.	The introduction was in a lecture theatre. Then they moved to two computer rooms which would have been familiar to the participants.
Participant sample	Same institution and school; Same year group of undergrads; Same gender make-up. All were students whose courses were largely creative.	Same institution and school; Same year group of undergrads; Same gender make-up. Their courses were less likely to contain a creative element.
Recruitment	Participants were approached individually and personally during workshop sessions. The task introduction was conducted on a personal basis.	Participants were recruited as a class during class time allocated by their lecturer. The task introduction was given to the class as a whole
Compensation/ motivation/ reward	Prior to starting the task participants were promised a "chocolate bar" as reward. On completion they were given their choice from a selection of 100g chocolate bars.	Participants were assigned the task as their work for that class, that day. They were given the opportunity to opt out by their lecturer; they all completed a consent form which permitted withdrawal.

Table C.12 - Comparing the feedback task conditions of pilot with main studies Underlined aspects are judged influential in the difference between the VAS results of the pilot main study.

The first three aspects (i.e. the *task material*, *workflow*, and *duration*) leading to repetitiveness and fatigue had been a concern during planning however this was discounted following the analysis of the readings over time during the task, as described in 10.3.1 and Figure 10.3.

"Physical location and setting", "*Recruitment"*, and "*Compensation/ motivation/ reward:* Being treated a) as part of a group and with perhaps some perceived element of compulsion (despite the opportunity to opt out) rather than b) as an individual, may well have engendered a less positive attitude among the main study participants. This may have affected their VAS judgments but there is no hard evidence for this.

However, the differences in the participant sample may be at the root of the differences in the VAS readings for the pilot and main studies. The different proportion of creative individuals within the samples may have had an effect. This difference was due to a) the main study participants being excluded from the pilot and b) time constraints and access to participants when recruiting for the pilot. Perhaps the pilot participants were behaving like the image-likers of the main study? Pilot participants were not asked about their format preferences, so there is no explicit basis for assigning them to the groups "image-liker" or "text-liker" as was done with the main study participants based on their post-task survey. However, if the pilot readings for Utility and Interest (raw, not log transformed), normalised 0-100, are compared with the corresponding readings from the main study (similarly normalised) similarities with the image-likers are seen. (Figure 10.10). To provide further evidence that the pilot participants are similar to the image-likers in the main study, and less similar to the text-likers, a Pearson Correlation analysis was done (figures in table below). The PCC for the Pilot VAS readings vs. the Image-liker VAS readings is 0.95 i.e. they are highly correlated (1.0 being a perfect 1:1 correlation). Whereas, comparing the same pilot VAS readings vs. the text-liker VAS readings, yields a PCC of 0.47 which is categorized as only a medium effect (Field 2009 p173).

Measure	Pilot	Image-likers, main	Text-likers, main
Utility_Text	23.233	34.856	25.374
Utility_Emotives	40.602	44.125	65.820
Utility_Abstracts	27.895	37.963	64.823
Interest_Text	40.301	49.230	55.803
Interest_Emotives	33.383	42.598	65.512
Interest_Abstracts	23.308	35.196	60.408

Table C.13 - The mean normalised (0 to 100) VAS readings used for Pearson Correlation analysis.The Pilot figures are from raw (not log transformed, readings).

This evidence suggests that the pilot participants, who were all studying a creative subject, were judging the Utility and Interest of the answer format in a similar way to the image-likers from the main study.

Appendix D Emotive SOM Construction

This Appendix accompanies Chapter 8.

Summary of Emotive Terms Survey

The survey was carried out by Kalkreuter (2013) and the returns handed to the author in a private communication. Following analysis the results were used to inform the selection of search terms used in the image screen scrape as described later.

18 subjects (staff and students) at TEX were asked to indicate on the Plutchik wheel which emotive terms they considered suitable for design feedback. 9 were categorised as designers (2 male); 9 were categorised as non-designers (4 male). Respondents either a) underlined a term (marking it as "most meaningful") b) left a term untouched (marking it as "meaningful") or c) crossed it out (marking it as "do not consider meaningful". A scanned example completed survey form can be found in the Additional Material directory, "Emotive SOM Construction".

The returned survey forms were coded thus

Response (and its meaning as per the survey instructions)	**Score**
Term scored out = Not meaningful	0
Term left untouched = Meaningful	1
Term underlined = Most meaningful	2

Table D.1 - Coding of design terms survey

The coding spread sheet is in Additional Material, "*Emotive SOM Construction*". The results and the image scrape search term selections are summarised in Table D.2. The selection was based on an analysis of the median and total scores accrued in the survey for each term. The comment column is used to explain any deviation from an even-handed approach to selection of the terms. The "Scrape weighting" column contains the value used to influence how many calls will be made to Google image search by the scrape script using that term and its synonyms.

Table D.3 shows the terms selected and rejected on the basis of that analysis. In terms of positive and negative emotions: on the full wheel 16 of the emotions can be categorised as negative and 16 as positive. In terms of the selected terms: of those proposed to be used in the scrape terms 6 are negative and 13 are positive.

Wheel spoke No.	Positive or negative emotion	Terms from the Plutchik wheel	Designers Most Meaningful (median = 2)	Non-Designers Most Meaningful (median = 2)	Designers Positive, i.e. both 1 and 2 taken to equal 1 (Median >0)	Non-Des Positive i.e. both 1 and 2 taken to equal 1 (Medn >0)	Designers Total>8 (9 subjects)	Non-Des Total >8 (9 subjects)	Comment	Include in search terms	Scrape Weighting
1	+	ecstasy				✓		✓			0
1	+	**joy**	✓	✓	✓	✓	✓	✓		✓	1
1	+	**serenity**	✓		✓	✓	✓	✓		✓	1
1a	+	**love**			✓	✓	✓	✓		✓	1
2	+	**admiration**	✓	✓	✓	✓	✓	✓		✓	1
2	+	**trust**	✓	✓	✓	✓	✓	✓		✓	1
2	+	**acceptance**	✓	✓	✓	✓	✓	✓		✓	1
2a	+	submission				✓					0
3	-	terror									0
3	-	fear			✓						0
3	-	**apprehension**	✓		✓		✓			✓	1
3a	+	**awe**			✓	✓	✓	✓		✓	1
4	+	**amazement**			✓	✓	✓	✓		✓	1
4	+	**surprise**		✓	✓	✓	✓	✓		✓	1
4	+	**distraction**	✓		✓	✓	✓	✓		✓	1
4a	-	**disapproval**			✓	✓			Included to bolster negative emotions	✓	1
5	-	grief									0
5	-	**sadness**	✓		✓	✓	✓			✓	1
5	-	**pensiveness**			✓	✓	✓			✓	1
5a	-	remorse				✓					0
6	-	loathing				✓					0
6	-	disgust			✓	✓					0
6	-	**boredom**			✓	✓	✓	✓		✓	1
6a	-	contempt				✓					0
7	-	rage									0
7	-	anger									0
7	-	annoyance									0
7a	-	**aggressiveness**			✓	✓	✓			✓	1
8	+	vigilance			✓	✓					0
8	+	**anticipation**			✓	✓	✓			✓	1
8	+	**interest**	✓	✓	✓	✓	✓	✓		✓	1
8a	+	**optimism**	✓		✓	✓	✓	✓		✓	1

Table D.2 - Analysis of the returns from the design terms survey, The "Wheel spoke no." column relates to the emotion family spokes on the Plutchik model. The "Include in search terms" column shows the conclusion of the analysis for a given term.

Selected Search terms (19)				**Rejected terms (13)**	
joy	acceptance	distraction	boredom	ecstasy	loathing
serenity	apprehension	disapproval	aggressiveness	submission	disgust
love	awe	sadness	anticipation	terror	contempt
admiration	amazement	pensiveness	interest	fear	rage anger
trust	surprise		optimism	grief	annoyance
				remorse	vigilance

Table D.3 - Terms selected and rejected from the model.

Search Terms for Image Screen Scrape

Search terms were formulated based on the 19 selected terms above and synonyms sourced from the MSWord Thesaurus. A database was constructed to facilitate the automation of the screen scrape. See Additional Material, "Emotive SOM Construction" folder, for scripts, database tables etc.

Gold Set Image Survey

20 participants were recruited, 10 (6 Male) at HWU campus and 10 (3 Male) at TEX. They undertook the task of tagging (on paper) 20 candidate quality control (QC) images as described below under heading, "Administering the Gold Set Image Questionnaire" (p.218). Typically, each spent around 12.5 minutes on task.

Administering the Gold Set Image Questionnaire

The package consisted of

For Participants:-

- A4 ring binder containing the images printed on A4 paper all from the same colour printer at the same time (to aid consistency of rendition).
- Each sheet was single sided was labelled with a letter as identifier, and placed in a polythene pocket for easy page turning.
- A form with 24 rows on which to indicate the image ID letter and details of the emotion(s) which the image evoked or depicted. The form also included the question "Is English your 1st language? Y/N".
- A version of the Plutchik emotion "wheel" with numbered spaces along with the emotion nouns. The colours were muted to allow clarity when reading the emotions and numbers.
- A sheet of dictionary definitions of all the nouns.
- An experiment participation agreement form and a pen.

Figure D.3 shows the kit as used by a participant.

For the administrator:-

- Introduction/instruction script to be followed while describing what the participants were to do and how.

- A version of the Plutchik wheel without numbers to aid in familiarisation with the model. (Specifically: 8 basic emotion spokes, 8 intermediate emotions, more intense emotions in the centre, less intense emotions to the outside.)
- The time each participant started and stopped to classify the images was noted.
- 120g of chocolate was given to each participant as a reward on completion.

All the materials referred to above can be found in the Additional Materials, "Gold Set image survey" folder.

Figure D.3 - Gold Set image survey kit as used by participants.

Gold Set Image Survey Results and Processing

The survey returns were entered and validated using spread sheets. QC emotion reference profiles (soon to be used to produce the Gold Set) were produced from the data. The spread sheets and MATLAB code for this are in the Additional Materials, "Emotive SOM Construction" folder.

Production of the Gold Set Data (QC Stimulus Patterns)

Vectors defining acceptable tags for the five Gold Set images were created from the QC emotion reference profiles. These Gold Set emotion pattern vectors, 1 for each QC stimulus, consisted of elements corresponding to each spot (1-56) on the emotion model (Figure 8.2). Each element was set to either 1 or 0 indicating a valid or invalid tag for that QC stimulus. They were constructed by accepting all the tags from the paper QC survey for the five Gold Set images, and adding two further acceptable tags on two of the images to fill in gaps on emotion family spokes, i.e. spot 25, 'serenity' on image 11799 and spot 31, 'distraction'.

The code, input, and output files for this are in the Additional Materials, "Emotive SOM Construction" folder.

The ECI Application

The ECI application allows users to tag images; delete tags; view the image in full view; read the help; and move on to the next image. At the end it displays a code to allow the participant to claim payment on their work provider site. Each participant sees the emotion model rotated at a random angle so as to prevent bias due to orientation. Two illustrative screens are shown below. A sequence of screens, including instructions screens, from a test run of the ECI are in the Additional Materials, "Emotive SOM Construction" directory.

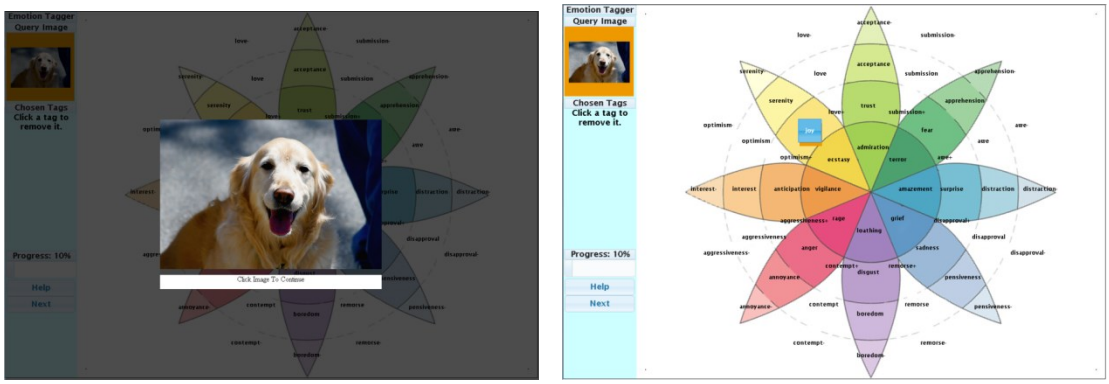


Figure D.4 - ECI Interface screens: Early in the task a stimulus is presented (left). Then the stimulus image is being dragged and dropped on 'joy' term spot (right).

ECI Experiment Database Manager App

Scripts used to manage the database tables are in a PHP application. See Additional Materials "Emotive SOM Construction" folder.

ECI Stimuli Packets

The application was tested with some volunteers to estimate the time required to classify an image. The aim was for a stimuli packet to take around 10-12 minutes. In deciding the number of stimuli per stimuli packet various factors were considered (See "Participant Pay" below).

The final decision was made to go with the following configuration totalling 32 stimuli with an expected typical time on task of around 10 minutes: Each sequence of stimuli to be tagged by participants consisted of 2 training stimuli followed by 25 actual stimuli interspersed with the five Gold Set images.

To manage the risk of the experiment not working, it was decided to divide the 2000 images into random batches of 100 images and generate batches of stimuli packets to allow batches of 100 to be completely classified before moving on to the next.

The 25 image stimuli in each stimuli packet were random within the batch (non-repeating within one stimuli packet) and balanced across the stimuli packets such that 1 batch of stimuli packets would produce 20 readings per image.

The stimuli packets were generated using MATLAB and spread sheets. These can be found in in the Additional Materials, "Emotive SOM Construction" folder.

Participant Pay

Factors taken into account were the quantity of images (2000), desired no of participant judgements per image (20); total cost to the project budget; fair pay a) in line with current worker expectations on CrowdFlower (Waterloo Unuversity 2013), at the time $0.50 for a 15 to 20 minute survey, and b) with reference to the UK minimum wage; suitably motivating pay but not too much to attract unethical workers (Kazai, 2011), estimated typical time on task, and participant fatigue.

It was decided to pay $1 per HIT (stimuli packet). This fitted with the level of pay in the earlier crowd task for the abstract image set.

Running the ECI App on CrowdFlower

The stimuli packets were posted as HITS on CrowdFlower. "Contributors" as CrowdFlower terms its workers saw the screen in the figure below, and could test their browser and screen resolution using the link in the HIT introduction before deciding to accept the HIT. (See Figure D.5)

A batch of stimuli packets was managed in this way:

1) HITs would be made available and this would give rise to
    a. Completed and satisfactory stimuli packets.
    b. Incomplete abandoned stimuli packets.
    c. Completed poor quality stimuli packets. (Described later).
2) Incomplete abandoned stimuli packets would be recycled.
3) Quality control assessment would be run on the completed stimuli packets assigning a quality score to the participant associated with the stimuli packet.

4) Any completed stimuli packets below a given QC threshold would be recycled.

5) Steps 1 to 4 would be repeated until no stimuli packets remain incomplete in the batch.

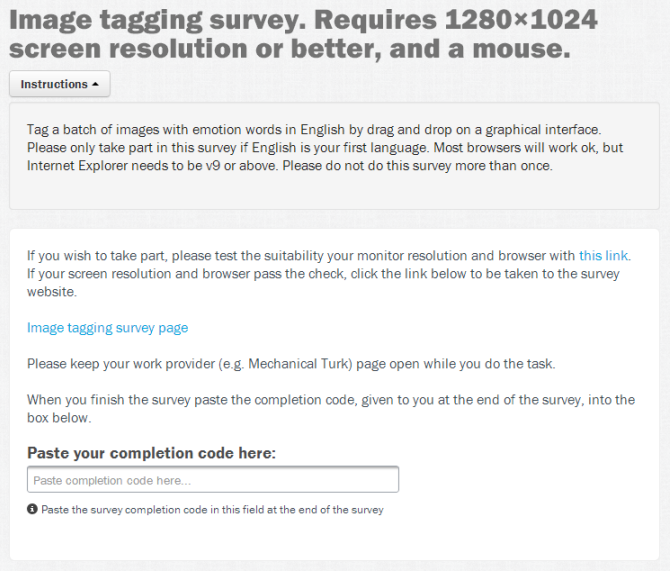


Figure D.5 - ECI HIT form as seen by CrowdFlower participants.

Each CrowdFlower job produced an output in the form of a CSV file detailing the claims made by participants against that job.

Commands in the ECI Experiment database management application allowed the tracking of participants and claims and the association of this data with the stimuli packets and the observations tables, by importing the claims data into the database. A properly formed claim would contain an identifier for the stimuli packet done by that participant, thus linking the claims table to the other tables in the database.

Assessing the Quality of the Crowdsourced Tags

The scoring described in 8.5.7 was achieved by a) extracting the observation data from the database with PHP/MySQL scripts b) processing this with MATLAB scripts to assign each set of observations a QC score and c) using this as input to further PHP/MySQL scripts to maintain a database of sets of observations and their QC scores (called the *Subjects* table) which linked to the actual observations (in the *Observations* table). This database allowed the extraction of all observations associated with sets of observations whose QC scores were over any given threshold.

The PHP/MySQL scripts for this are embedded as appropriately named commands in the ECI Experiment database management application. One of the ECI experiment DB

commands was named “Step 7 Process for QC”. This required running some MATLAB code taking a database report of the observations (output from a previous command) as input and producing a report assigning a QC score to a set of observations as output. The MATLAB output was then imported into the ECI experiment database into a table of participants (and their QC scores).

The MATLAB code compared the tags assigned by a participant for each of their five Gold Set images, with the Gold Set QC patterns. For each Gold Set image a tag matching the pattern scored 1 and a tag not matching the pattern scored 0. The mean tag score for each of the five Gold Set images was calculated. The QC score for that participant was the sum of all five mean Gold Set image tag scores. Thus a participant’s QC score can vary from zero to five. E.g. if a participant tagged all five Gold Set images with 2 tags each, and on the first four images both tags were correct (analogous to hits if the target is the pattern in the Gold Set data) but on the last Gold Set image one of the tags was wrong (or a miss) that participants score would be 4.5. (See example below; See also equation (8.1)).

$$\frac{1+1}{2}+\frac{1+1}{2}+\frac{1+1}{2}+\frac{1+1}{2}+\frac{1+0}{2}=4.5 \qquad (D.1)$$

(See Additional Materials, “Emotive SOM Construction” folder for MATLAB code and example inputs and output files).

Setting the Quality Control Threshold

As stated in Chapter 8, the ECI database consists of linked tables and allows the observations of individual participants to be sampled based on their QC score by running queries. When this was done it revealed that there were two reasons for a participant having a low QC score: 1) the obvious one of not sincerely attempting the task, but also 2) over tagging, by perhaps trying too hard and tagging with the maximum of five tags on each image. Such over taggers had misunderstood the instructions. However, whatever their motivation such over tagging would also produce suspect data with their tags not only including the valid tags but also other dubious tags. Thus by awarding low scores for both of these behaviours, the QC algorithm was doing its job.

As stated in Chapter 8, setting the threshold QC score at 3.1 and thus ruling out observations by participants scoring below that would safely prevent allowing the unreliable data from careless taggers and over enthusiastic taggers into the data set.

To put a QC score of 3.1 into context, tagging all five Gold Set images with a single good tag would give a score of 5.0 (i.e. 1.0 per Gold Set image) a score of 3.0 would be achieved by a participant completely miss-tagging two out of the five Gold Set images in their stimuli packet but properly tagging three of them. While a score of 3.1 requires that a participant perform reasonably well on 4/5 Gold Set images but still allows them to get one Gold Set image wrong. An over tagging participant might give good tags (analogous to hits) on all the Gold Set images but dilute their QC score by adding further incorrect tags (or misses).

Evaluating Effectiveness of Tagging in Early Batches

See Additional Materials, "Emotive SOM Construction", for MATLAB code and input/output files for generating the charts, SOM and dendrogram visualisations described in this subsection. This section refers to tag frequency vectors and also term vectors. These are described in 8.5.12 and also in *Assembling the Emotive2000 Emotion Profiles* here in Appendix D following this section. The three aids to evaluating the tagging are described blow.

a) Charts visualising the normalised tag frequencies for a given image were developed e. g. Figure 8.3.

b) SOM browsers were created. The Vesanto (1999) MATLAB SOM algorithm can be set to accept similarity matrixes or feature vectors. The tag frequency vectors were treated as feature vectors and thus used to inform the construction of SOM browsers which functioned just as the Abstract500 browser. In addition the final image thumbnails in the browsers linked to the database record for the image and its tag frequency chart.

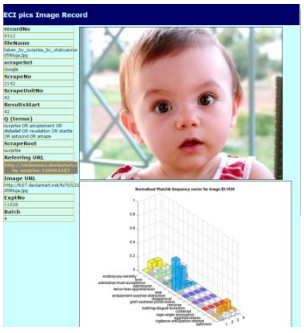


Figure D.6 - Screenshot of an image record in the ECI database. The record view shows the image and its tag frequency chart along with other data such as source URL and screen scrape search terms.

c) Using an interactive dendrogram application:

An interactive dendrogram web application (*The Dendrogrammer*) was built for the author's MSc project (code is provided in the Additional Materials). It allows visualisation of the output from MATLAB single linkage clustering. Scripts for clustering the data (based on the tagging category frequency vector for each image) were written and the clustering output fed to the *The Dendrogrammer*. Below is a figure showing one of the dendrogram views. The dendrogram was interactive in that clusters could be interrogated by clicking to reveal IDs of images. Part of the inputs to the *The Dendrogrammer* allows specification of a search application to which cluster data can be fed when a cluster is clicked on a dendrogram. This was set so as to call a query in the "ECI pics database manager" application thus displaying the images in the cluster.

See Additional Materials, "Emotive SOM Construction" for code and files used for the above processes.

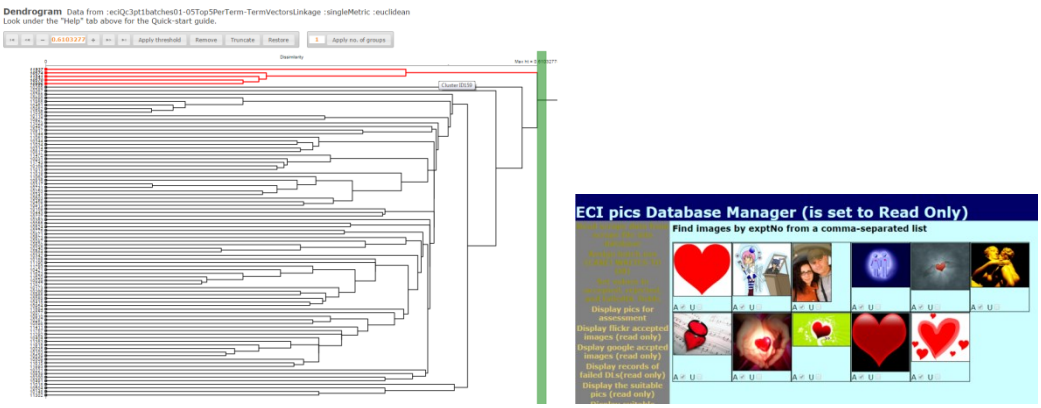


Figure D.7 - Screenshot of dendrogram (left) while assessing tagging in early batches .On the right is an image list query opened when clicking on a dendrogram cluster.

Assembling the Emotive2000 Emotion Profiles

After all 1600 stimuli packets had been completed by participants passing the QC threshold the ECI application was closed and data gathering ceased.

Scripts were run to output the quality controlled tagging observations. The output enabled creation of a spread sheet file, each row being a tagging observation which meets the QC threshold. These rows include these attributes: stimulus image ID; subject ID (the tagging participant from which the observation inherits its QC score); five tag values ranging from -1 for no tag to 55 the zero-indexed maximum emotion map tag location value.

While the ECI application prevented participants from tagging the same tag location on the emotion model twice, it was realised that it had permitted participants to tag different tag locations on the same term; e.g., on the tagging model, “love” appeared 3 times as “love-“, “love” and “love+”. Thus it was possible for of one term to be tagged three times by one participant. This occurred in only 303 observations out of over 40,000. However, as this would lead to some slight inconsistencies later when representing the data as normalised 32-term frequency vectors rather than simple tag frequency vectors it was decided to rectify this. This was achieved by locating all the affected observations within the quality controlled output and taking the more intense tag as that representing the participant’s reading of that term for that image (i.e. for the example of “love” being tagged three times the tag “love+” was accepted and the tags “love” and “love-“ were discarded). Tags for other terms in the same observation were not affected. This was done in MATLAB and the resulting output was a modified version the quality controlled tagging observations. The code which achieved this is in the Additional Materials, “Emotive SOM Construction” folder along with validation steps.

The processed and validated, quality controlled, tag observations were then further processed to produce, for each of the E2000 images, the normalised tag frequency vector and a chart to visualise it laid out on the emotion model. An additional view of the data based on the emotion terms rather than the location tags was also produced. (See chapter text).

See Additional Materials, “Emotive SOM Construction” for code and files used for the above processes.

The Emotive2000 Image Set in a SOM Browser

The Emotive2000 was viewed by assembling it in a SOM browser. As the 56-member tag frequency vectors represented the highest resolution data these vectors were used as feature vectors to inform SOM construction.

(The SOM along with the files and code to construct it and run it are in the Additional Materials, “Emotive SOM Construction”. It requires PHP enabled web space with a database connection to the E2000simplified MySQL database table.)

Filtering the Emotive2000 Image Set

The algorithm described in the figure below was developed to filter the Emotive2000 image set:

1	**Input**: ***Emotive 2000***
2	Set the desired number of images per term in the browser: ***TargetNo***.
3	For each of the design emotion term subset find the images whose highest frequency peak is that term: ***top-term images***.
4	By binary search find the minimum top-term frequency contrast (i.e. the smallest frequency gap between the top-term peak and the next nearest term peak for an image) that will satisfy *TargetNo*: ***MinContrast***.
5	Eliminate from the *top-term images* any images where the contrast between the peak term and the next highest peak is below *MinContrast*, leaving the ***top-term-high-contrast images.***
6	Sort these *top-term-high-contrast images* within terms by contrast.
7	For each term select the desired number of images per term from the *top-term-high-contrast images*, highest contrast first.
8	**Output**: **Filtered set e.g. *Emotive204***

Figure D.8 - Algorithm for filtering the Emotive2000 image set.

Two images were rejected (retrospective of assembling the Emotive2000); e.g. one had the word “optimism” printed in small font dead centre of the image. Therefore, a further input to the filter, a rejected images CSV list, was added.

An output from the above filtering, of 204 images, *Emotive204*, clustered into a SOM based on *tag frequency vectors* to make use of the full classification resolution on the images. See Additional Materials, “Emotive SOM Construction” for code and files used for this process.

Appendix E Evaluating Abstract500 & Summarisation

This Appendix accompanies Chapter 7.

The 20 feedback terms

Descriptive		Emotive	
Brittle	Flexible	Astonishment, surprise	Irritation, anger
Coarse	Smooth	Disgust, repulsion	Sadness, despair
Crumpling	Solid	Embarrassment, shame	Tenderness, feeling love
Delicate	Sticky	Enjoyment, pleasure	Wonderment, feeling awe
Fuzzy	Textured	Involvement, interest	Worry, fear

Table E.1 - The terms used in Task 1. Descriptive terms from Methven et al (2011) and emotive terms from Scherer (2005), specifically V. 2 of Geneva Emotion Wheel, in Sacharin et al (2012).

The Task 1 Interface

The Task 1 application was implemented in FlashBuilder4.6 and compiled as an iOS application which interfaced with a recording database using PHP. The figure below shows screens from the Task 1 interface.

The MATLAB code used to generate the Abstract500 SOM in the form of lines of MXML code for embedding in the application code prior to compilation can be found in the Additional Materials, "Evaluating Abstract500 & summarisation".

Figure E.1 - Screenshots from the Task 1 application SOM browser in 8x6 configuration (left) with a separate screenshot of an open stack (centre) and stimulus screen showing three participant selected images (right). After selecting three images participants tapped "Next" to save the selections and move on to the next stimulus. They could delete a chosen image and tap "Database" to return to the browser and select another image for the current stimulus.

Summarising the Task 1 Image Selections

A MATLAB script did all stages of the summarisation apart from rendering. It produced summary definition files which a PHP/JavaScript web application then rendered. The figure below shows screens from the Task 1 output viewer application.

See Additional Materials, “Evaluating Abstract500 & summarisation” for the files and code concerned.

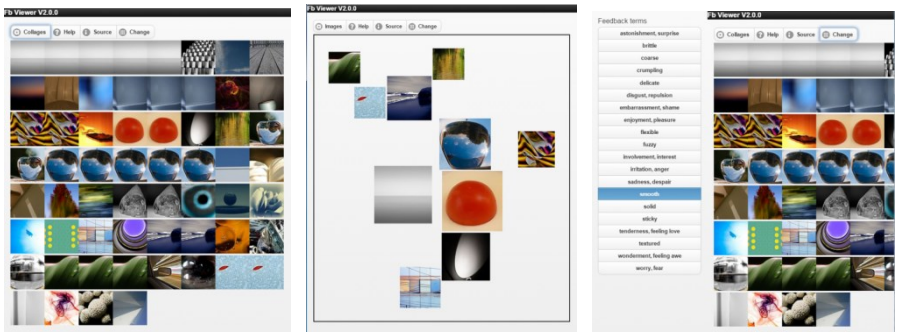


Figure E.2 - Screenshots from the Task 1 feedback viewer which allowed viewing of the output from Task 1. Shown are the image selection for “Smooth” (left), the summary for “smooth” (centre), a menu allowed selection to view output from other feedback terms (right).

Task 2 Interface and Recording Method

The interface was implemented in PHP, JavaScript (using jQuery), and MySQL. Stimuli packets generated in MATLAB and stored ready in a database.

Each participant would be served half of the stimuli (20 out of 40) in a random order, half being image lists and half being summaries (randomly) and spanning all 20 feedback terms. See Additional Materials, “Evaluating Abstract500 & summarisation” for the files and code concerned.

To manage the risk of technical failure during data gathering it was decided to produce the summaries beforehand, record them as screenshots in image files and serve the image files directly instead of as summary definition files to be rendered dynamically. It was simply one less thing to go wrong. There would also be a permanent record of the summary stimuli shown during the experiment. To allow more than one participant to be engaged in observations at any given time, access to the trials was controlled by unique trials login codes. In practice the author, as experiment admin, took care of setting up the iPads and logging into the trial before handing them over to a participant. This required logging on first on the master iPad, which automatically took on the role of master, and then logging on to the same trial with the slave which automatically ran in slave mode. Database fields keyed to that trial recorded the adoption of roles. The application was robust in that should a trial be interrupted through wifi connection loss it could be resumed from the last completed stimulus. Six iPads were carried allowing three sessions to be administered at any given time.

Steps were taken to vary the order of presentation of the 20 VAS items on the master display. (See Figure 7.6 and Figure 7.7). The feedback terms were always listed in two columns, one of descriptive and the other of emotive terms. Both these lists had a fixed order (alphabetical) but the last item on each list looped back to the first. Thus in effect no term was first on either list. Just the order was fixed. When a trial was logged into, the application established (at random) and then recorded a) placement of the descriptive terms on the left or the right and b) the point within both lists at which to list the terms from the top of their respective columns. If it became necessary to resume the session following interruption the database record meant that the positions were not re-randomised but were kept as per the first login for the trial.

Task 2 Detailed Results

Descriptive	*f*-1st		Emotive	*f*-1st	
Term	**List**	**Summary**	**Term**	**List**	**Summary**
brittle	0.167	0.167	astonishment, surprise	0.033	0.033
coarse	0.267	0.233	disgust, repulsion	0.033	0.133
crumpling	0.200	0.267	embarrassment, shame	0.000	0.100
delicate	0.167	0.333	enjoyment, pleasure	0.300	0.100
flexible	0.100	0.100	involvement, interest	0.133	0.033
fuzzy	0.200	0.333	irritation, anger	0.067	0.033
smooth	0.300	0.500	sadness, despair	0.200	0.267
solid	0.567	0.567	tenderness, feeling love	0.300	0.267
sticky	0.133	0.100	wonderment, feeling awe	0.267	0.167
textured	0.567	0.400	worry, fear	0.133	0.067

Table E.2 - f-1st for the 40 stimuli in Task 2.

Task 2 Normality Tests

In comparing groups of stimuli, *t*-tests were carried out on pairs of *f*-1st score distributions. The distributions were subjected to the Kolmogorov-Smirnov (K-S) test (Field, 2009 p144). This was done using SPSS. See the table below. For all 4 results distributions the significance value (Sig.) is not less than 0.05 indicating that none of them deviate significantly from normality (Field, 2009 p.246). It was inferred from this that parametric tests may be carried out on the distributions.

Distribution	K-S statistic	df	Sig.(*p*)	Passes test?
Comparing descriptive stimuli with emotive				
Descriptive	0.15	20	0.20*	Yes
Emotive	0.157	20	0.20*	Yes
Comparing image lists with summaries				
Lists	0.17	20	0.13	Yes
Summaries	0.16	20	0.19	Yes

Table E.3 - K-S tests for four results distributions where means were compared. 0.20 indicates that 0.20 is the lower bound of the true significance.*

Appendix F Constructing the Abstract SOM Image Browser

This Appendix accompanies Chapter 4.

Practical Parameters for the Image Screen Scrape

PP No	Practical Parameter	Rationale	Cf. ISR
1	Source from Flickr.	A brief exploration of the online service, Flickr (2015), showed that it would be a good source of abstract images and had a search facility allowing the specification of Creative Commons images only as search results. Flickr image records also contain an account name for the image owner to serve as attribution data. Google's (2015) image search service was also examined but as image attribution was more problematic (i.e. the only consistently available attribution data would be the image URL itself) it and abstract images were plentiful on Flickr was decided to use that service alone. Flickr also allows a resolution to be chosen.	ISR 7 Free; ISR 6 Resolution;
2	Gather 1800 images initially.	With the target number of acceptable images being 500 and the there being some categories to be excluded from the general category of abstract (ISR 2 and 3), in the first instance it was decided to gather 1800 images. This would allow at least 2/3 to be rejected. A number divisible by 30 was used as this was the default number of image records in a page of search results on Flickr	ISR 2 Non-specific; ISR3 No symbols; ISR 5 Population 500
3	Resolution 128x128 pixels min.	ISR 6 requires consideration of iPad screen resolution to allow deployment of the browser on that device. iPad1 is 1024x768 pixels. An image resolution of 128x128 would allow a SOM stack array of 8x6 at these resolutions. 8x6=48 stacks. 48 stacks containing 500 images would average at 10.4 images per stack. These SOM dimensions would be appropriate. It is likely that images would require reduced in size but, as the image type is to be abstract, loss of detail is not an issue. An image considered abstract reduced in size would not become less abstract by reduction.	ISR 6 Resolution; ISR 5 Population 500; ISRs 1, 2 and 3.
4	Safe search	Using a safe search for images will automatically rule out images with adult or offensive content. Such content would be ruled out by ISRs 1, 2 and 3 anyway so it would be best to filter these images out during the scrape so as not to waste time manually filtering them out later.	ISRs 1, 2 and 3.

Table F.1 - Practical parameters of the screen scrape for candidate images to populate the Abstract500 browser.

Candidate Images Accepted and Rejected in the Test Image Screen Scrape

The three tables below show the 20 images that were sampled from a test screen scrape and accepted or reject based on the Candidate Image Assessment Rules in Table 4.4.

Image	Reason for rejection	Image	Reason for rejection
	Not a full depiction of bikes. The depiction is an unconventional perspective. However, there is writing.		Full depiction of a nest.
	Has borders. Divided into two images i.e. not a single image.		Full depiction of robot or man figure. Also has borders
	People.		

Table F.2 - Five candidate images rejected after the test screen scrape, along with their reasons for rejection.

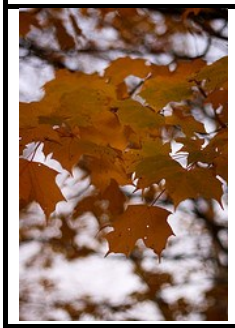
Image	Reservation Reason for Acceptance	Image	Reservation Reason for Acceptance
	A landscape. On close inspection it may be oil-painted Lack of definition renders it abstract.		A thicket Restricted view and monochrome produce a texture-like image.
	Leaves Restricted view with soft focus background. An unconventional perspective. Lack of definition renders it abstract.		A landscape An unconventional perspective

Table F.3 - Five borderline candidate images acceptedand reasons for acceptance.

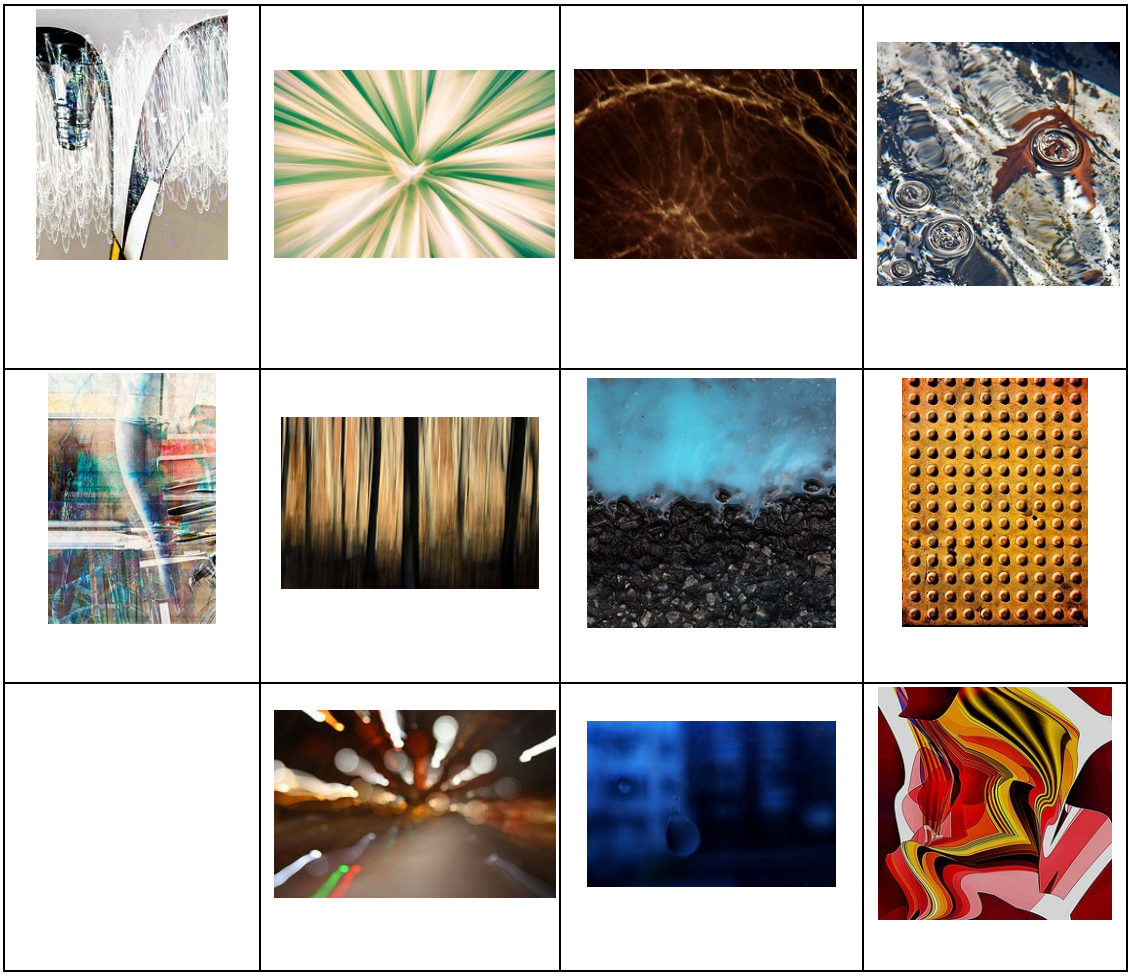


Table F.4 - Eleven candidate images accepted outright from the 20-image test scrape sample.

Resizing and Cropping to 128x128

This was done using *XnView* by batch processing.

The algorithm below describes the steps that were followed:

```
1   Make square:
2       If image is square then do nothing
3       Else crop left and right (or top and bottom) keeping the centre
4       End if
5   Resize:
6       Resize to 128x128
```

Figure F.3 - Algorithm for resizing and cropping the images following download.

Assessing Images for Suitability

Using the database facility for displaying, viewing and recording the assessment of the images in batches (described in Table 4.3, DBR No. 4) the images were assessed for

suitability against the criteria in Table 4.4 until 1010 had been assessed as suitable. The reason assessment was not ceased at 500 suitable images was to widen the spread of sample of images taken across a larger candidate population and dilute any bias which might be present in the order of the image scrape. (Precise figures: 1799 images were downloaded; 1515 were assessed; 505 were rejected; 1010 were flagged as acceptable; 284 were left not assessed.) Thus the rejection rate for images was 505/1515 = 33%.

Elimination of Duplicate Images

The possibility of duplicate images existed. This was addressed by using MATLAB to calculate the mean RGB values for each image, storing the RGB figures in three fields (r,g,b) of the database. The three dimensional RGB colour space was divided up systematically and, using a form to input red, green, and blue, the database was queried to display the images in screens of up to 60 images at a time sorting on the red figure for checking. Duplicate images would be expected to appear side-by-side when displaying the query results. This process revealed one instance of duplication within the suitable images and those two images were marked "unsuitable" to rule them out of the final set.

The Approach to Quality Control

A script was written to triage the completed stimuli packets, flag those which qualified for bonus payment, and highlight for scrutiny those where the participant time on task was below a time threshold set at four minutes. The script implemented the algorithm in the figure below.

1	Set *SCRUTINISE*, *BONUS*, and *COMPLETED* flags to false
2	Calculate *Average No. of Likenesses* per image and *time on task*
3	If *time on task* < 4 minutes then set *SCRUTINISE* to true
4	End if
5	If *Average No. of Likenesses* > 2.5 then set *BONUS* to true
6	End if
7	If *No. of completed queries* =20 then set *COMPLETED* to true
8	End if

Figure F.4 - Algorithm for triage of completed stimuli packets.

Stimuli packets flagged *COMPLETED* = false were recycled to be done again. Those flagged *SCRUTINISE* = true were scrutinised (see below). Those which passed scrutiny were a) accepted into the data b) the participant was paid and c) if flagged *BONUS*=true the participant was additionally paid the bonus.

To enable scrutiny of stimuli packets flagged *SCRUTINISE* = true, a MATLAB script was created to take a completed stimuli packet as input and display it as 20 columns of images. (A stimuli packet consisted of 20 query images and resulted in 20 corresponding likeness lists each of two to four images). Each column consisted of the query image at the top and the two to four likenesses provided by the participant below it. This window was stretched across two large displays such that it could be viewed in a single view. Rather than just subjectively second-guessing the participant's judgements, account was taken of the bootstrap SOM layout and whether or not a) the participant had drilled down into the structure seeking likenesses b) had ranged across the SOM stacks or c) had merely accessed image stacks nearest to the "Next" button or grabbed the top image of each stack. Plausible results sets from those marked *SCRUTINISE* = true, were accepted, otherwise they were rejected and no payment made. The figure below shows an example output from the script enabling scrutiny of a single stimuli packet results set.

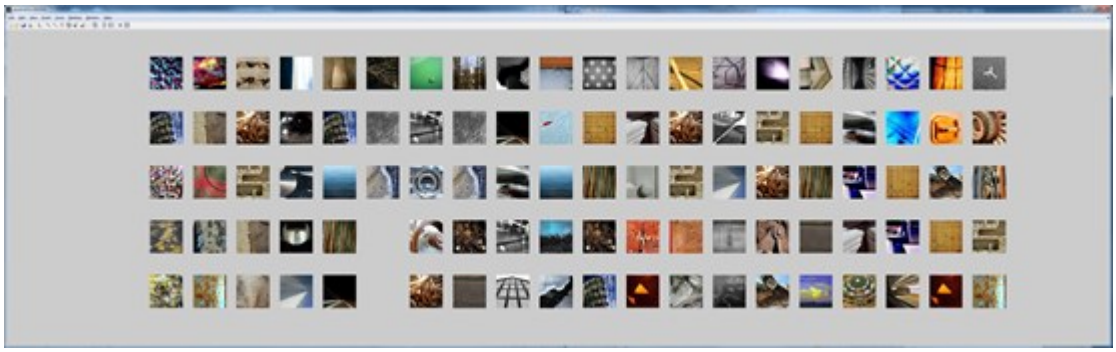


Figure F.1 - Example output from the script enabling scrutiny of stimuli packets. Each column is one observation. 20 columns equates to all of a single participants observations. A column's top image is the query image. The two to four images below that are the participant's likeness selections from the bootstrap browser.

It would be possible to create an algorithm based on the position within the bootstrap SOM browser of the likeness images to calculate an estimate of minimum browsing effort required to generate each likeness list and use that as a basis on which to accept or reject hurriedly produced observations. However, as there were only 200 stimuli packets to be done and the percentage requiring scrutiny was not great the scrutiny was done manually and the cost of developing such an algorithm was avoided.

The Bootstrap Sort

The apparatus consisted of a table with a large white surface, a swivel chair, and the 100 bootstrap subset images printed on white paper in colour at 128x128 resolution. Each printed image was presented on a playing card sized piece of paper with its ID number at the bottom to ensure consistent orientation and a barcode encoding that ID on the back for swift and accurate recording of the data after each sorting session. (See figure below.)

20 participants (11 male) were recruited being invited to attend the lab or the studio (depending on the campus) and offered 100g chocolate as reward. The mean time on task was 17.6 minutes (median: 17; SD: 4.8; max.: 28; min.: 9).

The participants were instructed on how to carry out the sort following the steps set out by Halley (2012) and reported in Padilla et al. (2013). This meant they could sort the images into as many groups as they wished the only provisos being that they must deem the images in each group to be similar and that any singleton image must form its own group of one (i.e. an aggregated group of singletons was disallowed).

After each sorting session a participant's groups were recorded in a spread sheet which a) enabled data entry using a bar code reader and b) contained formulae not only to create formatted output but also to do a reconciliation checks to validate data entry.

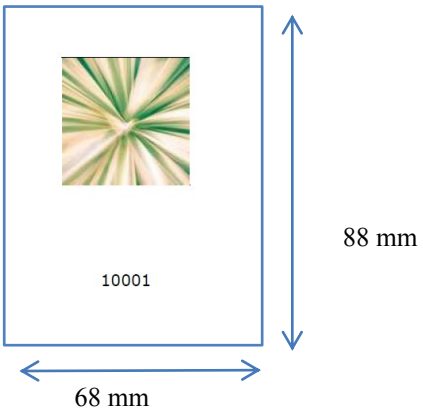


Figure F.2 - Layout of one of the free sort image cards created for the bootstrapping.

The Crowdsourced Enrichment of the Matrix

The target for the augmentation was 10 presentations each of all 400 query (or augmentation) images. As in Halley's (2012) procedure, each stimuli packet was one AMT HIT (human intelligence task (Kazai, 2011)). HITs consisted of 20 query images

from the 400. Thus 400 query images occurring 10 times would make 4000 queries/20 per HIT meant that 200 stimuli packets would be required. 200 such packets representing a balanced but random spread of the query images were produced. However, due to the nature of the HIT flow through the augmentation application it would be necessary to have a small number of additional packets to allow an orderly termination of the AMT HIT batches. (This was due to the nature of assessment of HIT results which can result in a proportion of the results being rejected and thus their associated stimuli packets being recycled through the augmentation application again until 200 HITs with accepted results had been achieved). Therefore, an extra 20 stimuli packets were created by duplicating a random sample of 20 from the original 200. This meant that there could not be certainty about the number of presentations of each query image but between 9 and 11 times was expected. (See the table summarising opportunities below the details about payment.)

The same pay as offered by Halley (2012) was offered (see figure below).

IMPORTANT: Payment Criteria and Experiment Consent

1) Minimum payment: We have ways of checking that the similarity judgments you give during the experiment are valid. Please do not continue if you intend to rely on random chance as the threshold is much higher. Only a valid set of judgements will qualify for the payment of $0.75.
2) Bonus: We encourage you to take the experiment seriously give your best judgements about which images are most similar to the query image. Indeed, a bonus payment of $0.25 will be paid if you exceed an average of 2.5 selections per query image rather than the minimum of 2, while still making careful judgements.
3) We will keep an anonymous copy of the judgements you make.
4) We may decide to publish the results of our experiment.
5) You may withdraw from the experiment at any time.
6) By continuing you are consenting to the above.

Finally, please remember not to use your web browser BACK button or REFRESH during the experiment. Just use the buttons in the experiment until after you are given your claim token at the end.

Thank you for taking part.

Figure F.3 - Wording of the payment Criteria and Consent dialog in the augmentation application as used on AMT for the Abstract 500 augmentation.

Stimuli packet results were checked according to the algorithm on page 234 of this Appendix. Incomplete and rejected sets of observations, were recycled until all 200 of the desired stimuli packets had a completed acceptable set of results associated with them. To identify stimuli packets for recycling they were assessed in batches. The table below illustrates the statistics on incompleteness, rejection, and bonus qualification.

Description	%
Incomplete and recycled	9
Rejected and recycled.	2.5
Accepted and standard payment paid.	13.5
Accepted and paid with bonus.	75

Table 11.1 - Statistics from assessing a typical batch of stimuli packet results for completeness, rejection, and bonus payment.

After 200 acceptable sets of results were collected the application was removed from AMT and the data were processed. That processing involved the intermediate step of calculating an opportunities (or presentations) matrix with which to normalise the final similarity matrix. That opportunities matrix was also used to survey the number of presentations to allow an overview of the frequency with which query images were presented. The table below sets out the number of presentations of the query images.

No. of Opportunities (or Presentations)	Frequency	Frequency x Opportunities for Reconciliation
7	2	14
8	15	120
9	84	756
10	179	1790
11	120	1320
Total for Reconciliation		4000

Table F.5 - Table summarising the opportunities (or presentations) of the 400 query images. It shows the frequency with which the number of opportunities (which ranged from 7 to 11) occurred. The reconciliation shows how this was achieved within the 200 stimuli packets (200x20=4000 queries; an average of 10 per image).

The bootstrap SOM used for the augmentation application can be found in the Additional Materials, “Constructing the Abstract500 SOM browser” folder.

The first stage of processing the output from the augmentation application was likeness vectors for each of the 400 augmentation images. See example in the Table F.6. These likeness vectors (one for each of the 400 augmentation, or query, images) and the 100x100 bootstrap similarity matrix were input to code adapted from exemplar code from Halley (2011). This generated the new 500x500 augmented similarity matrix using

the method described in Halley (2012) and reported in Padilla et al. (2013). This incrementally adds the new augmentation (or query) images to the matrix. Each time a new image is added it is assigned a similarity vector representing the mean similarity values of the images selected by the augmentation participants as being most similar to that new image (i.e. those likened to it). The resulting similarity matrix creates a convincing organisation for the image set. See the chapter text.

Augmentation image ID (a query image)	Likeness Vector
101	94,18,75,56,55,86,83,51,71,17,2,86,32,2,70,22, 67,32,85,71,18,83,18

Table F.6 - An example of a likeness vector produced during processing of the output from the augmentation application. Each member of the vector identifies a bootstrap image which a participant likened to the query image, ID 101, when choosing the 2 to 4 images from the bootstrap browser they judged most similar to image 101. Repetitions are likely in the likeness vectors as the 10 participants (on average) viewing a query image often agree.

Evaluating the Perceptual Data Using MDS

A 3D visualisation of the Abstract500 (created as described in the chapter text) showed clear regions and themed clusters. (Figure F.4).

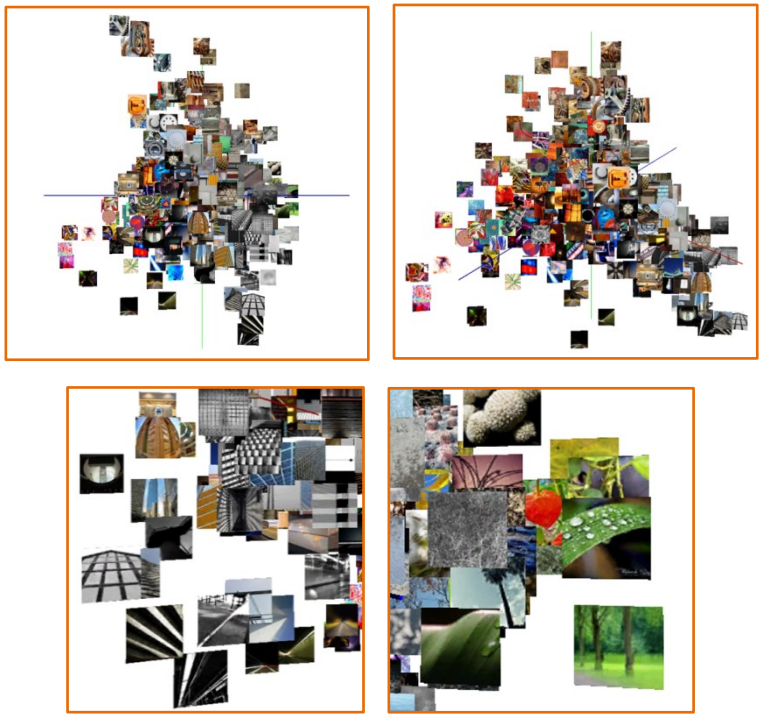


Figure F.4 - Classical MDS 3D view. Screenshots of two further aspects from the view, shown in Figure 4.4 (top). Two clusters, one structural themed, another natural themed (bottom).

Appendix G Summary of Experimental Sessions

Experimental session	Chapter ref	Participants
Free grouping 100 abstract images for bootstrap browser	4.5.5	20
Task 1 – Terms-to-Images	7.3.1	20
Task 2 – Images-to-Terms	7.7.1	60
Gold set image survey	8.5.3	20
Pilot evaluation feedback task	9.7.1	10
Total		**130**

Table G.1 - Face-to-face sessions.

Interview session	Chapter ref	Participants
Pilot evaluation designer interview	9.9.3	1
Main evaluation designer interviews	10.4.3	12
Total		**13**

Table G.2 - Interviews.

En Bloc Experimental session	Chapter ref	Participants
Main evaluation feedback task	10.3.1	**32**

Table G.3 - En bloc session

Crowdsourced experiment	Chapter ref	Accepted results sets
Abstract500 matrix augmentation (AMT)	4.5.6	200
ECI application (CrowdFlower)	8.5.11	1600
Total		**1800**

Table G.4 - Crowdsourced sessions.

Questionnaires completed	Chapter ref	Participants
Design emotion terms survey	Page 216	18
Main evaluation post task survey	10.3.5	31
Total		**49**

Table G.5 - Questionnaires.

Session type	Sets of data collected and analysed
Face-to-face experiment, interview or en bloc	175
Questionnaires completed	49
Crowdsourced accepted results sets	1800
Total	**2024**

Table G.6 - Summary: Total sets of human task, interview, or questionnaire data collected and analysed.

Bibliography

Ahern, S., King, S., Naaman, M., & Nair, R. Summarization of Online Image Collections via Implicit Feedback. *WWW'07:ACM International Conference on World Wide Web*, 1325-1326.

Amazon Mechanical Turk (2015), https://www.mturk.com (Accessed 24th Jan 2015).

Aristotle. (1962). *Politics*. (T. A. Sinclair, Trans.). Penguin. (Original work circa 350 BC.)

Ashby, F.G., Queller, S., & Berretty, P.M. (1999). On the dominance of unidimensional rules in unsupervised categorization. *Perception & Psychophysics*, 61 (6), 1178-1199.

Baranovskiy, D. (2010). *Raphaël JavaScript Library*. http://raphaeljs.com (Accessed 12th Jan 2015).

Brabham, D.C. (2012). The myth of amateur crowds: A critical discourse analysis of crowdsourcing coverage. *Information, Communication & Society*, 15(3), 394-410.

Bradley, M.M., Codispoti, M., Cuthbert, B.N., & Lang, P.J. (2001). Emotion and motivation I: defensive and appetitive reactions in picture processing. *Emotion*, 1 (3), 276.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine, *Computer Networks and ISDN Systems*, 30(1–7), 107-117.

Bringer, J.D., Johnston, L.H., & Brackenridge, C.H. (2006). Using computer-assisted qualitative data analysis software to develop a grounded theory project. *Field Methods*, 18 (3), 245-266.

Cambria, E., Livingstone, A., & Hussain, A. (2012). The hourglass of emotions. *Cognitive Behavioural Systems*, Springer. 144-157.

Chandler, D. (2002). *Semiotics : the basics (2nd ed.)*, Routledge.

Chen, C., Gagaudakis, G., & Rosin, P. (2000). Similarity-Based Image Browsing. *XVI IFIP World Computer Congress, International Conference on Intelligent Information Processing*, 206–213.

Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., & Vedaldi, A. (2014). Describing Textures in the Wild. *CVPR 2014: IEEE Conference on Computer Vision and Pattern Recognition*, 3606-3613.

Clarke, A.D., Halley, F., Newell, A.J., Griffin, L.D., & Chantler, M.J. (2011) Perceptual similarity: a texture challenge. *BMVC2011: 22nd British Machine Vision Conference*.

Coffield, F., Moseley, D., Hall, E., & Ecclestone, K. (2004). *Should we be using learning styles?: what research has to say to practice.* Learning & Skills Research Centre.

Cook, E., Teasley, S.D., & Ackerman, M.S. (2009) Contribution, commercialization & audience: understanding participation in an online creative community. *Group2009: ACM International Conference on Supporting Group Work*, 49-50.

Corbin, J., & Strauss, A. (2008). *Basics of qualitative research: Techniques and procedures for developing grounded theory*. Sage.

Cox, T.F., & Cox, M.A.A. (2001). *Multidimensional Scaling (Second Edition).* Chapman & Hall/CRC.

Craft, B.C., P. (2006). Using Sketching to Aid the Collaborative Design of Information Visualisation Software-A Case Study. *Human Work Interaction Design: Designing for Human Work ,* Springer, 221, 103-122.

CrowdFlower (2015), http://www.crowdflower.com (Accessed 23rd Feb 2015).

Culler, J. (1976). *Saussure*, Fontana.

Dan-Glauser, E., & Scherer, K. (2011). The Geneva affective picture database (GAPED): a new 730-picture database focusing on valence and normative significance. *Behavior Research Methods*, 43 (2), 468-477.

Darwin, C. (1965). *The expression of the emotions in man and animals.* University of Chicago Press. (Original work 1872).

Davis, A. (2009). New media and fat democracy: The paradox of online participation. *New media & society*. 11 (7), 1083 - 1264.

Depalov, D., Pappas, T.N., Li, D., & Gandhi, B. (2006). Perceptually based techniques for semantic image classification and retrieval. *SPIE Human Vision and Electronic Imaging XI*, 6057.

Delplanque, S., N'diaye, K., Scherer, K., & Grandjean, D. (2007). Spatial frequencies or emotional effects?: A systematic measure of spatial frequencies for IAPS pictures by a discrete wavelet analysis. *Journal of Neuroscience Methods*, 165 (1), 144-150

Dribbble (2015). http://dribbble.com (Accessed 23rd Feb 2015).

Donne, J. (1975). *Devotions upon emergent occasions*: Oxford University Press. (Original work 1623) (Quote from Meditation, chapter XVII).

Dow, S., Gerber, E., & Wong, A. (2013) A pilot study of using crowds in the classroom. *CHI'13: ACM Conference on Human Factors in Computing Systems*, 227-236.

Eckert, C., & Stacey, M. (2000). Sources of inspiration: a language of design. *Design Studies*, 21(5), 523-538. doi: 10.1016/S0142-694X(00)00022-3

Egorova, M., Safonov, I., & Korobkov, N. (2008). Collage for Cover of PhotoBook . *GRAPHICON'2008: International Conference on Computer Graphics and Vision,* 160-163.

Ekman, P. (1984). Expression and the nature of emotion. *Approaches to emotion*. LEA. 319-344

Ekman, P. (1999). Basic emotions. *Handbook of cognition and emotion*, 98, 45-60.

Entropia Partners (2015) http://www.entropiapartners.com (Accessed 24th Feb 2015).

Facebook (2015). https://www.facebook.com (Accessed 24th Feb 2015).

Epstein, S. (1994). Integration of the cognitive and the psychodynamic unconscious. *American Psychologist*, 49 (8), 709.

Estelles-Arolas, E., & Gonzalez-Ladron-de-Guevara, F. (2012). Towards an integrated crowdsourcing definition. *Journal of Information Science*, 38 (2), 189-200.

Evans, J.S.B. (2003). In two minds: dual-process accounts of reasoning. *Trends in cognitive sciences*, 7 (10), 454-459.

Evans, J.S.B. (2008). Dual-Processing Accounts of Reasoning, Judgment, and Social Cognition. *Annual Review of Psychology*, 59 (1), 255-278

Evans, J.S.B. (2013). Two minds rationality. *Thinking & Reasoning*, 20 (2), 129-146.

Evans, J.S.B., & Stanovich, K.E. (2013). Dual-process theories of higher cognition advancing the debate. *Perspectives on Psychological Science*, 8 (3), 223-241.

Everitt, B. (1974). *Cluster Analysis*. Heinemann.

Fan, J., Gao, Y., Luo, H., Keim, D.A., & Li, Z. (2008) A novel approach to enable semantic and visual image summarization for exploratory image search. *MIR'08: ACM International Conference on Multimedia Information Retrieval*, 358-365.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods,* 39 (2), 175-191.

Favorskaya, M., Yaroslavtzeva, E., & Levtin, K. (2012). Intelligent Collage System. *IIMSS'12: 5th International Conference on Intelligent Interactive Multimedia Systems and Services*, Springer, 341-350. DOI 10.1007/978-3-642-29934-6.

Fendrich, K., & Hoffmann, W. (2007). More than just aging societies: the demographic change has an impact on actual numbers of patients. *Journal of Public Health,* 15 (5), 345-351.

Field, A. (2009). *Discovering Statistics Using SPSS (3rd ed.).* Sage.

Flickr (2015) https://www.flickr.com (Accessed 24th Feb 2015).

Galton, F. (1907a). Vox Populi (The Wisdom of Crowds). *Nature*, 75, 450-451.

Galton, F. (1907b). One vote, one value. *Nature*, 75, 414.

Garcia, I. (2013). Learning a Language for Free While Translating the Web. Does Duolingo Work? *International Journal of English Linguistics*, 3(1), p19.

Garner, S., & McDonagh-Philp, D. (2001). Problem interpretation and resolution via visual stimuli: the use of 'mood boards' in design education. *Journal of Art & Design Education*, 20(1), 57-64.

Gevins, A., & Smith, M.E. (2000). Neurophysiological measures of working memory and individual differences in cognitive ability and cognitive style. *Cerebral Cortex*, 10 (9), 829-839.

Google (2011), "A fall spring-clean", Google Official Blog September 2, 2011 http://googleblog.blogspot.co.uk/2011/09/fall-spring-clean.html (Accessed 3rd May 2015).

Google (2015) https://www.google.co.uk (Accessed 24th Feb 2015).

Guiraud, P. (1971). *Semiology*. Routledge.

Halley, F. (2011). Code exemplars: SOM browser construction and similarity matrix augmentation. Private communication to the author.

Halley, F. (2012). *Perceptually Relevant Browsing Environments for Large Texture Databases*, PhD thesis, Heriot-Watt University, Edinburgh.

Hariri, A.R., Mattay, V.S., Tessitore, A., Fera, F., & Weinberger, D.R. (2003). Neocortical modulation of the amygdala response to fearful stimuli. *Biological Psychiatry*, 53 (6), 494-501.

Hebecker, R., & Ebbert, C. (2010). Creation and Validation of Symbols with Purposeful Games and Online Survey. *DESIRE'10: ACM Network Conference on Creativity and Innovation in Design*, 112-120.

Heesch, D. (2008). A survey of browsing models for content based image retrieval. *Multimedia Tools and Applications*, 40 (2), 261-284.

Hill, R.A., & Dunbar, R.I.M. (2003). Social network size in humans. *Human Nature*, 14 (1), 53-72.

Hofmans, J., & Theuns, P. (2008). On the linearity of predefined and self-anchoring Visual Analogue Scales. *British Journal of Mathematical and Statistical Psychology*, 61 (Pt 2).

Horton, J.J., & Chilton, L.B. (2010) The labor economics of paid crowdsourcing. In *Proceedings of the 11th ACM conference on Electronic commerce*, pp 209-218.

Hsu, W.H., Kennedy, L.S., & Chang, S.-F. (2007). Video search reranking through random walk over document-level context graph. *MM'07: The 15th ACM International Conference on Multimedia*, 971-980.

Instagram (2015) https://instagram.com/ (Accessed 24th Feb 2015).

Izard, C.E. (1971). *The face of emotion*. Appleton-Century-Crofts.

Jakobson, R. (1960). Closing statement: Linguistics and poetics. *Style in language*, 350, 377.

Junghöfer, M., Bradley, M.M., Elbert, T.R., & Lang, P.J. (2001). Fleeting images: A new look at early emotion discrimination. *Psychophysiology*, 38 (2), 175-178.

Kajiyama, T., & S., S.i. (2014) An Application Search Interface Including Sense-related Search Facets. *ICMR '14: ACM International Conference on Multimedia Retrieval*, 463.

Kalkreuter, B. (2013). Emotive Terms Survey Returns. Private communication to the author.

Kalkreuter, B., & Robb, D. (2012). HeadCrowd: Visual feedback for design. *Nordic Textile Journal, Sustainability and Innovation in the Fashion Field* (1), 70-81.

Kalkreuter, B., Robb, D., Padilla, S., &Chantler, M.J. (2013), Managing creative conversations between designers and consumers, in Britt, H., Wade, S., Walton, K. (Eds), *Futurescan 2: Collective Voices*, Association of Fashion and Textile Courses, Sheffield, 90-99.

Kalos, M.H., & Whitlock, P.A. (2009). *Monte Carlo Methods*, Wiley-VCH.

Kaplan, A.M., & Haenlein, M. (2011). Two hearts in three-quarter time: How to waltz the social media/viral marketing dance. *Business Horizons*, 54 (3), 253-263.

Kazai, G. (2011). In search of quality in crowdsourcing for search engine evaluation. *Advances in information retrieval,* Springer ,165-176.

Kerminen, P., & Gabbouj, M. (1999). Image Retrieval Based on Color Matching. *FINSIG'99: IEEE Finnish Signal Processing Symposium*, 89-93.

Kerminen, P., & Gabbouj, M. (2000). Visual Goodness Evaluation of Color-based Retrieval Processes. *EUSIPCO 2000: European Signal Processing Conference*. 2153-2156.

Kerminen, P., Tanttu, J.T., & Gabbouj, M. (2003). Utilization of luminance information on color-based image retrieval. *Nordic MATLAB Conference*, Comsol A/S, 252 - 256.

Keil, A., Bradley, M.M., Hauk, O., Rockstroh, B., Elbert, T., & Lang, P.J. (2002). Large-scale neural correlates of affective picture processing. *Psychophysiology*, 39 (5), 641-649.

Kirman, B., Lawson, S., Linehan, C., Martino, F., Gamberini, L., & Gaggioli, A. (2010) Improving social game engagement on facebook through enhanced socio-contextual information. *CHI'10: ACM Conference on Human Factors in Computing Systems,* 1753-1756.

Kirman, B., Lineham, C., & Lawson, S. (2012). Exploring mischief and mayhem in social computing or: how we learned to stop worrying and love the trolls. *CHI'12 Extended Abstracts on Human Factors in Computing Systems*, 121-130.

Klein, D., & Manning, C.D. (2003) Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics* - Volume 1

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, 78 (9), 1464-1480.

Kohonen, T. (1998). The self-organizing map. *Neurocomputing*, 21 (1), 1-6.

Kosara, R., & Ziemkiewicz, C. (2010) Do Mechanical Turks dream of square pie charts? *ACM 3rd BELIV'10 Workshop: BEyond time and errors: novel evaLuation methods for Information Visualization,* 63-70.

Kozhevnikov, M. (2007). Cognitive styles in the context of modern psychology: Toward an integrated framework of cognitive style. *Psychological Bulletin*, 133 (3).

Kruskal, J.B. (1964a). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29 (1), 1-27.

Kruskal, J.B. (1964b). Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, 29 (2), 115-129.

Kvale, S., & Brinkmann, S. (2009). *Interviews: Learning the craft of qualitative research interviewing (Second ed.)*, Sage.

Lang, P., & Bradley, M.M. (2007). The International Affective Picture System (IAPS) in the study of emotion and attention. *Handbook of emotion elicitation and assessment*, 29.

Lee, M.H., Singhal, N., Sungdae, C., & Park, I.K. (2010). Mobile photo collage. *CVPRW'10: IEEE Computer Vision and Pattern Recognition Workshops*, 24-30.

Lee, S.-S., & Yong, H.-S. (2008). Ontosonomy: Ontology-based extension of folksonomy. *IWSCA'08: IEEE International Workshop on Semantic Computing and Applications*, 27-32.

Legland, D. (2009). *geom3d*: MATLAB CENTRAL File Exchange. http://www.mathworks.com/matlabcentral/fileexchange/24484-geom3d/content/geom3d/geom3d/, (Accessed 23rd Feb 2015).

Lerner, J.S., Small, D.A., & Loewenstein, G. (2004). Research Report Heart Strings and Purse Strings Carryover Effects of Emotions on Economic Decisions. *Psychological Science*, 15 (5), 337-341.

Lim, Y.-k., Donaldson, J., Jung, H., Kunz, B., Royer, D., Ramalingam, S., Stolterman, E. (2008). Emotional experience and interaction design, *Affect and Emotion in Human-Computer Interaction.* Springer , 116-129.

Lowe, D.G. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2), 91-110.

Luo, J., Joshi, D., Yu, J., & Gallagher, A. (2011). Geotagging in multimedia and computer vision—a survey. *Multimedia Tools and Applications*, 51(1), 187-211.

Luther, K., Pavel, A., Wu, W., Tolentino, J.-l., Agrawala, M., Hartmann, B., & Dow, S.P. (2014). CrowdCrit: crowdsourcing and aggregating visual design critique. *CSCW2014: ACM Conference on Computer Supported Cooperative Work & Social Computing, Companion*, 21-24.

Maclennan, B., Kypri, K., Langley, J., & Room, R. (2012). Non-response bias in a community survey of drinking, alcohol-related experiences and public opinion on alcohol policy. *Drug and Alcohol Dependence*, 126 (1-2), 189-194.

Marchewka, A., Żurawski, Ł., Jednoróg, K., & Grabowska, A. (2014). The Nencki Affective Picture System (NAPS): Introduction to a novel, standardized, wide-range, high-quality, realistic picture database. *Behavior Research Methods*, 46 (2), 596-610.

McCandless, D. (2009). *Information is beautiful*. Collins.

McCrum-Gardner, E. (2008). Which is the correct statistical test to use? *British Journal of Oral and Maxillofacial Surgery*, 46 (1), 38-41

Martinez, W.L., Martinez, A., & Solka, J. (2011). *Exploratory data analysis with MATLAB Second Edition*, CRC Press.

McMinn, A.J., Moshfeghi, Y., & Jose, J.M. (2013). Building a large-scale corpus for evaluating event detection on twitter. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*. 409-418.

Meagher, M.W., Arnau, R.C., & Rhudy, J.L. (2001). Pain and Emotion: Effects of Affective Picture Modulation. *Psychosomatic Medicine*, 63 (1), 79-90.

Mei, T., Rui, Y., Li, S., & Tian, Q. (2014). Multimedia search reranking: A literature survey. *ACM Computing Surveys*, 46(3), 1-38

Methven, T.S., Orzechowski, P.M., Chantler, M.J., Baurley, S., & Atkinson, D. (2011) A Comparison of Crowd-Sourcing vs. Traditional Techniques for Deriving Consumer Terms. In *Digital Engagement '11*, Newcastle, UK.

Mikels, J.A., Fredrickson, B.L., Larkin, G.R., Lindberg, C.M., Maglio, S.J., & Reuter-Lorenz, P.A. (2005). Emotional category data on images from the International Affective Picture System. *Behavior research methods*, 37 (4), 626-630.

Mikels, J.A., Löckenhoff, C.E., Maglio, S.J., Carstensen, L.L., Goldstein, M.K., & Garber, A. (2010). Following your heart or your head: focusing on emotions versus information differentially influences the decisions of younger and older adults. *Journal of Experimental Psychology: Applied*, 16 (1), 87.

Mikels, J.A., Maglio, S.J., Reed, A.E., & Kaplowitz, L.J. (2011). Should I go with my gut? Investigating the benefits of emotion-focused decision making. *Emotion*, 11 (4), 743.

Mizerski, R.W., & White, J.D. (1986). Understanding and using emotions in advertising. *Journal of Consumer Marketing*, 3 (4), 57-69.

Nazir, A., Raza, S., & Chuah, C.-N. (2008). Unveiling facebook: a measurement study of social network based applications. *IMC '08: ACM SIGCOMM Conference on Internet Measurement,* 43-56.

Nagamachi, M. (1995). Kansei engineering: a new ergonomic consumer-oriented technology for product development. *International Journal of industrial ergonomics*, 15 (1), 3-11.

Nederhof, A.J. (1985). Methods of coping with social desirability bias: A review. *European Journal of Social Psychology,* 15 (3), 263-280.

Neurath, O. (1936). *International Picture Language*. Kegan Paul, Trench & Trubner Co.

Nisbett, R.E., Peng, K., Choi, I., & Norenzayan, A. (2001). Culture and systems of thought: holistic versus analytic cognition. *Psychological review*, 108 (2), 291.

Norman, D.A., & Ortony, A. (2003). Designers and users: Two perspectives on emotion and design. Paper presented at the Proc. of the Symposium on Foundations of Interaction Design at the Interaction Design Institute, Ivrea, Italy.

Norman, G., Sherbino, J., Dore, K., Wood, T., Young, M., Gaissmaier, W., Monteiro, S. (2014). The etiology of diagnostic errors: A controlled trial of system 1 versus system 2 reasoning. *Academic Medicine*, 89 (2), 277-284

Padilla, S. (2011). *Task-1 experiment application code*. Private communication to the author.

Padilla, S., Robb, D., Halley, F., & Chantler, M.J. (2012). Browsing Abstract Art by Appearance. *Predicting Perceptions: Proceedings of the 3rd International Conference on Appearance*, 100-103.

Padilla, S., Halley, F., Robb, D., & Chantler, M., (2013), Intuitive Large Image Database Browsing Using Perceptual Similarity Enriched by Crowds, *Computer Analysis of Images and Patterns, LNCS*, Springer, 8048, 169-176.

Pammi, S., & Schroder, M. (2009). Annotating meaning of listener vocalizations for speech synthesis. *ACII 2009*: *3rd International Conference on Affective Computing and Intelligent Interaction and Workshops Proceedings*. pp 3304 – 3309.

Park, I.K., Yun, I.D., & Lee, S.U. (1999). Color image retrieval using hybrid graph representation. *Image and Vision Computing*, 17(7), 465-474.

Piller, F., Schubert, P., Koch, M., & Möslein, K. (2005). Overcoming Mass Confusion: Collaborative Customer Co-Design in Online Communities. *Journal of Computer-Mediated Communication*, 10 (4).

Plato. (1998). *Republic (Vol. 237)*. (Waterfield, R., Trans.). Oxford University Press. (Original work *circa* 380 BC.)

Plutchik, R. (1997). The circumplex as a general model of the structure of emotions and personality. *Circumplex models of personality and emotions*, R. P. H. R. Conte (Ed.), APA, 17-45.

Plutchik, R. (2003). *Emotions and life: perspectives from psychology, biology, and evolution*. APA.

Plutchik, R., & Conte, H.R. (1997). *Circumplex models of personality and emotions*. APA.

Porter, C.E., Devaraj, S., & Sun, D. (2013). A Test of Two Models of Value Creation in Virtual Communities. *Journal of Management Information Systems*, 30 (1), 261-292.

Power, M.J. (2006). The structure of emotion: An empirical comparison of six models. *Cognition & Emotion*, 20 (5), 694-713.

Randall, D.M., & Fernandes, M.F. (1991). The social desirability response bias in ethics research. *Journal of Business Ethics*, 10 (11), 805-817.

Rayner, S., & Riding, R. (1997). Towards a categorisation of cognitive styles and learning styles. *Educational Psychology*, 17 (1-2), 5-27.

Reddit (2015). http:// http://www.reddit.com/ (Accessed 23rd Feb 2015).

Reips, U., & Funke, F. (2008). Interval-level measurement with visual analogue scales in Internet-based research: VAS Generator. *Behavior Research Methods*, 40 (3), 699-704.

Riding, R.J. (1997). On the nature of cognitive style. Educational Psychology, 17(1-2), 29-49.

Riding, R.J., & Ashmore, J. (1980). Verbaliser-imager learning style and children's recall of information presented in pictorial versus written form. *Educational Studies*, 6, 141-145.

Riding, R., & Cheema, I. (1991). Cognitive Styles—an overview and integration. *Educational Psychology*, 11(3-4), 193-215..

Robb, D.A., Padilla, S., Kalkreuter, B., & Chantler, M.J. (2015a). Crowdsourced Feedback With Imagery Rather Than Text: Would Designers Use It? *CHI2015: ACM Conference on Human Factors in Computing Systems,* 1355-1364.

Robb, D.A., Padilla, S., Kalkreuter, B., & Chantler, M.J. (2015b). Moodsource: Enabling Perceptual and Emotional Feedback from Crowds. *CSCW2015: ACM Conference on Computer Supported Cooperative Work & Social Computing, Companion*, 21-24.

Rogowitz, B.E., Frese, T., Smith, J.R., Bouman, C.A., & Kalin, E. (1998.). Perceptual image similarity experiments. *SPIE Conference on Human Vision and Electronic Imaging*, 576-590.

Romero, D., & Molina, A. (2011). Collaborative networked organisations and customer communities: value co-creation and co-innovation in the networking era. *Production Planning & Control*, 22 (5-6), 447-472

Ross, R.T. (1938). A statistic for circular scales. *Journal of Educational Psychology*, 29 (5), 384-389.

Rother, C., Bordeaux, L., Hamadi, Y., & Blake, A. (2006). AutoCollage. SIGGRAPH 2006: ACM Transactions on Graphics, 25 (3), 847-852.

Russell, J.A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, 39 (6), 1161-1178.

Sacharin, V., Schlegel, K., & Scherer, K.R. (2012*). Geneva Emotion Wheel rating study* (Report). University of Geneva, Swiss Center for Affective Sciences.

Sanders, E., & Simons, G. (2009). A Social Vision for Value Co-creation in Design. *Open Source Business Resource (Value Co-Creation)*.

Sanders, E., & Westerlund, B. (2011) Experiencing, Exploring, and Experimenting in and with Co-Design Spaces. *Nordes 2001*: *Nordic Design Research Conference.* Helsinki.

Savage, N. (2012). Gaining wisdom from crowds. *Communication*. ACM, 55(3).

Scherer, K.R. (2005). What are emotions? And how can they be measured? *Social science information*, 44 (4), 695-729.

Schmidt, F.A. (2013) The Good, The Bad and the Ugly: Why Crowdsourcing Needs Ethics. *CGC2013, Third International Conference on Cloud and Green Computing.* pp 531-535

Schwarz, N., Bless, H., & Bohner, G. (1991). Mood and persuasion: Affective states influence the processing of persuasive communications. *Advances in experimental social psychology*, 24, 161-199.

Sharma, N., Rawat, P., & Singh, J. (2011). Efficient CBIR Using Color Histogram Processing. *Signal & Image Processing : An International Journal(SIPIJ)*, 2(1).

Schroeder, M., & Noy, P. (2001). Multi-agent visualisation based on multivariate data. *AGENTS '01: ACM International Conference on Autonomous Agents,* 85-91.

Siegert, I., Bock, R., Vlasenko, B., Philippou-Hubner, D., & Wendemuth, A. (2011). Appropriate emotional labelling of non-acted speech using basic emotions, geneva emotion wheel and self assessment manikins. *ICME'11: IEEE International Conference on Multimedia and Expo*.

Silva, V.D., & Tenenbaum, J.B. (2002). Global versus local methods in nonlinear dimensionality reduction. In *Proceedings Advances in neural information processing systems*.

Silverman, D. (2010). *Doing qualitative research: A practical handbook (Third ed.),* SAGE.

Sloman, S.A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, 119 (1), 3-22.

Smeulders, A.W., Worring, M., Santini, S., Gupta, A., & Jain, R. (2000). Content-based image retrieval at the end of the early years. *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* 22 (12), 1349-1380.

Soleymani, M., & Pantic, M. (2012). Human-centered implicit tagging: Overview and perspectives. *SMC 2012:* IEEE International Conference on Systems, Man, and Cybernetics, 3304 – 3309.

Surowiecki, J. (2004). *The wisdom of crowds: why the many are smarter than the few and how collective wisdom shapes business, Economies, Societies and Nations.* Anchor.

Swan, A.R.H., & Sandilands, M. (1995). *Introduction to geological data analysis*. Blackwell. 174-177.

Talwar, A., Jurca, R., & Faltings, B. (2007). Understanding user behavior in online feedback reporting. *EC'07: ACM Conference on Electronic Commerce*. 134-142.

Tan, L., Song, Y., Liu, S., & Xie, L. (2011). ImageHive: Interactive Content-Aware Image Summarization. *IEEE Computer Graphics and Applications*, 32 (1), 46-55.

Taylor, R.K. (2000). Marketing strategies: gaining a competitive advantage through the use of emotion. *Competitiveness Review: An International Business Journal incorporating Journal of Global Competitiveness*, 10 (2), 146-152.

Tenenbaum, J.B., de Silva, V., & Langford, J.C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, 290, 2391-2323.

Tiedens, L.Z., & Linton, S. (2001). Judgment under emotional certainty and uncertainty: the effects of specific emotions on information processing. *Journal of personality and social psychology*, 81 (6), 973.

Toffler, A. (1980). *The Third Wave*. William Collins Sons and Co. Ltd.

Tourangeau, R. (2004). Survey research and societal change. *Annual Review of Psychology*, 55, 775-801.

Tsytsarau, M., & Palpanas, T. (2012). Survey on mining subjective data on the web. *Data Mining and Knowledge Discovery*, 24 (3), 478-514.

Tuzovic, S. (2010). Frequent (flier) frustration and the dark side of word-of-web: exploring online dysfunctional behavior in online feedback forums. *Journal of Services Marketing*, 24 (6), 446-457.

Twitter (2015) https://twitter.com/ (Accessed 4th May 2015).

University of Waterloo (2013). *Human Participant Research Guidelines. Use of Crowdsourcing Services*, University of Waterloo, Ontario, Canada. https://uwaterloo.ca/research/sites/ca.research/files/uploads/files/crowdsourcing_guidelines_access_check_done.pdf (Accessed 14/4/2015)

Vesanto, J., Himberg, J., Alhoniemi, E., & Parhankangas, J. (1999) Self-organizing map in Matlab: The SOM Toolbox. *Matlab DSP Conference*. 35-40.

Von Ahn, L., & Dabbish, L (2004). Labeling images with a computer game, *CHI2004: ACM Conference on Human factors in computing systems*, pp 319-326.

Von Ahn, L., & Dabbish, L. (2008). Designing Games With A Purpose. *Communications of the ACM*, 51(8), 58-67.

Wang, F., & Kan, M.-Y. (2006). NPIC: Hierarchical synthetic image classification using image search and generic features, *Image and Video Retrieval* Springer, pp. 473-482.

Wallin, L. (2010). *2D Games Engine*. http://www.lukewallin.co.uk (Accessed 24th Feb 2015).

Wang, J., Sun, J., Quan, L., Tang, X., & Shum, H.-Y. (2006). Picture Collage. *CVPR'06: IEEE International Conference on Computer Vision and Pattern Recognition*, 347-354.

Watson, D., & Clark, L.A. (1992). Affects separable and inseparable: On the hierarchical arrangement of the negative affects. *Journal of Personality and Social Psychology*, 62 (3), 489-505.

Webster, J., & Ho, H. (1997). Audience engagement in multimedia presentations. *ACM SIGMIS Database*, 28 (2), 63-77.

Witteman, C., van den Bercken, J., Claes, L., & Godoy, A. (2009). Assessing Rational and Intuitive Thinking Styles. *European Journal of Psychological Assessment*, 25 (1), 39-47

Wittenburg, K., Ali-Ahmad, W., LaLiberte, D., & Lanning, T. (1998). Rapid-fire image previews for information navigation. *AVI'98: ACM International Working Conference on Advanced Visual Interfaces*. 76-82.

Xu, A., Huang, S.-W., & Bailey, B.P. (2014). Voyant: Generating Structured Feedback on Visual Designs Using a Crowd of Non-Experts. *CSCW2014: ACM Conference on Computer Supported Cooperative Work & Social Computing*, 37-40.

Xu, A., Rao, H., Dow, S.P., & Bailey, B.P. (2015). A Classroom Study of Using Crowd Feedback in the Iterative Design Process.*CSCW2015: ACM Conference on Computer Supported Cooperative Work & Social Computing,* (To Appear).

Xu, H., Wang, J., Hua, X.-S., & Li, S. (2011). Hybrid image summarization. *MM'11: ACM International Conference on Multimedia*. 1217-1220.

YouTube (2015) https://www.youtube.com (Accessed 24th Feb 2015).

Yuki, M., Maddux, W.W., & Masuda, T. (2007). Are the windows to the soul the same in the East and West? Cultural differences in using the eyes and mouth as cues to recognize emotions in Japan and the United States. *Journal of Experimental Social Psychology*, 43 (2), 303-311.